NORDEUROPÄISCHE ARBEITEN ZUR LITERATUR, SPRACHE UND KULTUR / NORTHERN EUROPEAN STUDIES IN LITERATURE, LANGUAGE AND CULTURE 17

Maria Håkansson Ramberg

Validität und schriftliche Sprachkompetenz

Eine Studie zur Bewertung schriftlicher Leistungen im Fach Deutsch an schwedischen Schulen



NORDEUROPÄISCHE ARBEITEN ZUR LITERATUR, SPRACHE UND KULTUR / NORTHERN EUROPEAN STUDIES IN LITERATURE, LANGUAGE AND CULTURE 17

Maria Håkansson Ramberg

Validität und schriftliche Sprachkompetenz

Die Validität bei der Bewertung sprachlicher Kompetenzen im Fach Deutsch als Fremdsprache steht im Mittelpunkt der vorliegenden Studie. Untersucht wird, auf welche sprachlichen, textuellen und aufgabenbezogenen Aspekte die Bewertenden Wert legen, die Bewerterübereinstimmung schwedischer Bewertender und wie sich schwedische Bewertungen schriftlicher Kompetenz zu den Niveaus des Gemeinsamen europäischen Referenzrahmens (GER) verhalten. Neben einer qualitativen thematischen Inhaltsanalyse bietet die Arbeit auch quantitativ angelegte statistische Berechnungen.
Auf der Grundlage der Ergebnisse werden didaktische Empfehlungen für den schulischen Fremdsprachenunterricht gegeben, die auch über den schwedischen Kontext hinaus Bedeutung haben.
Dieses Buch wendet sich an Forschende und Praktiker im fremdsprachlichen Bereich, an Lernende, Lehrkräfte, Lehrwerkautorinnen und -autoren, Lehramtsstudierende sowie weitere Akteure schulischer Bildung.

Die Autorin

Maria Håkansson Ramberg absolvierte ein Lehramtsstudium der Fächer Deutsch, Schwedisch und Schwedisch als Fremdsprache und verfügt über eine langjährige Lehrerfahrung im schwedischen Bildungssystem. Sie studierte Germanistik, Skandinavistik, Literaturwissenschaft und Sprachdidaktik in Lund, Freiburg im Breisgau und Växjö.
Maria Håkansson Ramberg promovierte am Institut für moderne Sprachen der Universität Uppsala. Ihre Interessenschwerpunkte liegen in den Bereichen Sprachlehr- und -lernforschung, Bewertung sprachlich-kommunikativer Kompetenzen sowie in der Implementierung des Gemeinsamen europäischen Referenzrahmens in Schweden.

www.peterlang.com

Validität und schriftliche Sprachkompetenz

NORDEUROPÄISCHE ARBEITEN ZUR LITERATUR, SPRACHE UND KULTUR /

NORTHERN EUROPEAN STUDIES IN LITERATURE, LANGUAGE AND CULTURE

Herausgegeben von / Edited by Frank Thomas Grub

BAND / VOLUME 17

Zu Qualitätssicherung und Peer Review der vorliegenden Publikation

Die Qualität der in dieser Reihe erscheinenden Arbeiten wird vor der Publikation durch externe, von der Herausgeberschaft benannte Gutachter im Double Blind Verfahren geprüft. Dabei ist der Autor der Arbeit den Gutachtern während der Prüfung namentlich nicht bekannt; die Gutachter bleiben anonym.

Notes on the quality assurance and peer review of this publication

Prior to publication, the quality of the work published in this series is double blind reviewed by external referees appointed by the editorship. The referees are not aware of the author's name when performing the review; the referee's names are not disclosed.

Maria Håkansson Ramberg

Validität und schriftliche Sprachkompetenz

Eine Studie zur Bewertung schriftlicher Leistungen im Fach Deutsch an schwedischen Schulen



PETER LANG

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.d-nb.de abrufbar.

Diese Publikation ist dank der Unterstützung des Instituts für moderne Sprachen der Universität Uppsala entstanden.

ISSN 2196-9760
ISBN 978-3-631-87372-4 (Print)
E-ISBN 978-3-631-88890-2 (E-PDF)
E-ISBN 978-3-631-88891-9 (EPUB)
DOI 10.3726/b20146





Peter Lang – Berlin · Bern · Bruxelles · Istanbul · New York · Oxford · Warszawa · Wien

Diese Publikation wurde begutachtet.

www.peterlang.com

Abstract: Validity is a fundamental aspect of quality within the field of testing and assessment. Despite this fact, little research has been conducted on the validity of written assessment in a Swedish educational context and even less regarding assessment in an additional foreign language. In particular, there is little work on teachers' scoring and the relation between ratings of students' performances in upper secondary education and the external reference levels of the CEFR, the widely used framework from Council of Europe.

Against this background, the present study was designed with the aim of examining the validity of the assessment of students' written language proficiency in German at different steps according to the Swedish school system. The focus is on (1) raters' construct conceptualisation, (2) inter-rater consistency of the Swedish raters, and (3) the relationship between Swedish ratings and ratings at a B1 level of the CEFR. The student data comprise 60 texts written in L2 German by Swedish upper secondary school students in courses aiming for three different steps of the national curriculum. The essays were scored by (a) students' own teachers, b) external Swedish raters using Swedish national performance standards and (c) external CEFR raters in terms of the B1 level of the CEFR. Analysis of scores and rater comments were based on qualitative and quantitative methods, for example qualitative data analysis (QDA) and descriptive, correlational and reliability statistics.

The results were analysed in relation to theoretical concepts of validity and central validation frameworks. The findings show that raters pay attention to a wide range of aspects in students' written proficiency, although, to some extent, different interpretations of how student texts should be evaluated in relation to the national performance standards and a tendency to comment more on certain aspects could be observed. Analyses of inter-rater consistency indicate that the ability of Swedish raters to rank students' performances is satisfactory, but that there are challenges in reaching general agreement, especially for intermediate and higher scores. Additionally, the results suggest that a pass level of *Tyska 5* at upper secondary school is roughly equivalent to a B1-level of the CEFR. Finally, the thesis highlights the importance of rater training and discussions about assessment as part of strengthening teachers' professional assessment competence. By exploring validity from different perspectives, the study provides a more complete picture of learner written assessment in an additional foreign language in Sweden and contributes to a deepened conceptual understanding of validity aspects within a Swedish educational context.

Keywords: L2 writing assessment, validity, modern foreign language, L2 German, Swedish upper secondary school, Common European Framework of Reference for Languages (CEFR), language education

Inhaltsverzeichnis

Abbildungs- und Tabellenverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Vorwort

Die vorliegende Publikation ist eine überarbeitete Fassung meiner Doktorarbeit, die im Fach Deutsch an der Universität Uppsala entstand. Zum Entstehen dieser Arbeit haben mehrere Personen auf unterschiedliche Weise beigetragen.

Mein besonderer Dank gilt meinem Betreuer Prof. Dr. Frank Thomas Grub (Uppsala) sowie meiner Zweitbetreuerin Prof. Gudrun Erickson (Göteborg), die meinen Forschungsprozess mit wertvollen Vorschlägen und großer Fachkompetenz engagiert unterstützt haben.

Prof. Dr. Monika Angela Budde (Vechta), Prof. Dr. Eva Breindl (Erlangen-Nürnberg) und Prof. Dr. Ute Bohnacker (Uppsala) haben frühere Versionen der Arbeit gelesen und konstruktive Kritik und Ratschläge gegeben. Dr. Andrea Meixner (Uppsala/Berlin) hat das Manuskript sorgfältig Korrektur gelesen. Bei den Schülerinnen und Schülern, Lehrkräften und Schulleitern an den teilnehmenden Schulen sowie bei den externen Bewertenden bedanke ich mich für ihre Kooperationsbereitschaft.

Für die finanzielle Unterstützung zweier Bewertender und die Bereitstellung eines Übungssatzes für Jugendliche (das Modul *Schreiben* des Goethe-Zertifikates B1) möchte ich dem Goethe-Institut meinen Dank aussprechen, besonders Dr. Katharina Buck (Stockholm/Kiew) und Stefanie Dengler (München). Nicht zuletzt ermöglichte das Institut für moderne Sprachen der Universität Uppsala durch seine finanzielle Unterstützung den Druck der vorliegenden Publikation.

Beim *Peter Lang Verlag* bedanke ich mich ganz herzlich für die freundliche Betreuung.

Der größte Dank geht an meine Familie: meinen Mann David, der mir in allen Promotionsphasen liebevoll und unterstützend mit klugen Ratschlägen zur Seite gestanden hat, sowie unsere beiden Kinder Erik und Hedvig dafür, dass Ihr mit eurer Lebensfreude und neugieriger Sicht auf die Welt zeigt, was im Leben wichtig ist.

Uppsala, im Dezember 2022
Maria Håkansson Ramberg

1. Einleitung

Spätestens im Zuge der Digitalisierung verlieren zahlreiche Staats-, Landes- und weitere Grenzen an Bedeutung. Dies führt zu neuen sprachlichen Herausforderungen und einem großen Bedarf an Sprachkompetenzen, sowohl in Englisch als auch in weiteren Sprachen. Gleichzeitig wird die Bedeutung fremdsprachlicher Kompetenzen in der heutigen Zeit auch in verschiedenen Richtlinien und sprachpolitischen Dokumenten betont (z. B. Europäische Union 2014; Skolverket 2018a; Council of Europe 2020). Wenn das Ziel der Europäischen Kommission, dass alle Menschen in Europa in zwei Fremdsprachen neben der eigenen kommunizieren können (vgl. European Council 2002), erreicht werden soll, ist ein auf kommunikative Kompetenzen ausgerichteter Fremdsprachenunterricht in den jeweiligen Bildungssystemen der europäischen Länder unabdingbar. Hierbei wird meist das im Jahr 2001 vom Europarat publizierte Dokument *Gemeinsamer europäischer Referenzrahmen für Sprachen: lernen, lehren, beurteilen*[1] (Europarat 2001, im Folgenden abgekürzt als GER oder Referenzrahmen) als Grundlage und Bezugspunkt für Sprachenlernen, Sprachunterricht und die Bewertung[2] von Sprachkompetenzen verwendet. Der GER verfolgt einen grundsätzlich handlungsorientierten Ansatz, wonach Menschen als sozial Handelnde angesehen werden, die kommunikative Aufgaben bewältigen können sollten. Der Referenzrahmen hat demnach ein kompetenzorientiertes Verständnis von Sprachverwendung, wobei insbesondere die kommunikative Sprachkompetenz hervorgehoben wird (vgl. Europarat 2001: 21).

Der kompetenzorientierte Fremdsprachenunterricht hat in den letzten Jahrzehnten zunehmend an Bedeutung gewonnen. Der kommunikative Ansatz, der sich bereits in den schwedischen Bildungsstandards für die Fremdsprachen aus den 80er Jahren zum Ausdruck kommt, hat in schwedischem Kontext eine lange Tradition. Anders als im Fall von Englisch begegnen schwedische

1 Das Originaldokument, *The Common European Framework of Reference for Languages: learning, teaching, assessment* (CEFR), ist im selben Jahr auch auf Deutsch erschienen. In der vorliegenden Arbeit wird vorwiegend die oben erwähnte deutschsprachige Ausgabe aus dem Jahr 2001 verwendet.

2 Die Begriffe *Bewertung*, *Beurteilung* und *Evaluation* bzw. *bewerten*, *beurteilen* und *evaluieren* werden im Weiteren als Synonyme verwendet.

Schülerinnen und Schüler aber den sog. modernen Sprachen (*Moderna språk*)[3] fast ausschließlich in einem schulischen Kontext. Dies bedeutet, dass schwedische Lehrkräfte in diesem Schulfach eine tragende Rolle für die aktive Sprachverwendung und das Erlernen der Sprache haben. Schwedische Lehrkräfte haben zudem im Vergleich zu ihren Kolleginnen und Kollegen in vielen anderen Ländern eine größere Autonomie bei der Gestaltung des Unterrichts, tragen aber auch eine vergleichsweise hohe Verantwortung für die Bewertung von Kompetenzen der eigenen Schülerinnen und Schüler (vgl. Nusche et al. 2011). Angesichts dessen müssen schwedische Lehrkräfte nicht nur über die erforderlichen Fachkenntnisse verfügen, sondern auch eine Vertrautheit mit und ein Verständnis von Zielen, Methoden, Rahmenbedingungen und Prozessabläufen sowie ihren Konsequenzen hinsichtlich der Bewertung haben.

Die Bewertung sprachlicher Kompetenz gehört zu den zentralen Aufgaben der Lehrkräfte und ist eines der wichtigen Elemente, um die Qualität schulischer Bildung zu sichern, wie auch der Untertitel des GER zeigt. Eine *valide* und zuverlässige Bewertung ist aber eine Voraussetzung dafür, dass Testergebnisse mit Legitimität im und außerhalb des schulischen Kontextes verwendet werden können. Die Ergebnisse einer Bewertung der zweiten Fremdsprache im schwedischen System können auch im Hinblick auf den Zugang zu weiteren Studien oder bestimmten Berufsbranchen eine große Bedeutung für die Lernenden haben, sog. *High-Stakes-Prüfungen*[4]. Die Bewertung fremdsprachlicher Kompetenz in einem schwedischen Schulkontext erfolgt nach einem System mit Wissensanforderungen, wobei versucht wird, die Gesamtkompetenz der Schülerinnen und Schüler in einem Urteil zu erfassen. Diese geschieht häufig durch verschiedene Teiltests der sprachlichen Kompetenz, wie das Prüfen des Lese- und Hörverstehens sowie der mündlichen und schriftlichen Interaktion und Produktion. Bei einer Bewertung freier Sprachverwendung, d. h. mündlicher und schriftlicher Sprachkompetenz, ist allerdings schwer zu vermeiden, dass eine gewisse Subjektivität bei der Bewertung eine Rolle spielt.

Auch wenn sich aktuell ein erhöhtes Interesse an Bewertung und Fragen der *Validität* sowie Gleichwertigkeit im Bildungsbereich in Schweden abzeichnet, gibt es bislang verhältnismäßig wenige wissenschaftliche Arbeiten, die sich mit

---

3 In Schweden ist Englisch die erste Fremdsprache. In Klasse 6 in der Grundschule kann eine zweite Fremdsprache aus dem Spektrum der als *Moderna språk* („moderne Sprachen") bezeichneten Sprachen gewählt werden.

4 Ein *High-Stakes-Test* ist, im Gegensatz zu einem *Low-Stakes-Test*, für die Testteilnehmenden von großer Bedeutung, indem das Testergebnis verwendet wird, um wichtige zukünftige Entscheidungen zu treffen (z. B. Zugang zu einer höheren Ausbildung).

Fragestellungen hinsichtlich der Bewertung und der Validität in einer zweiten Fremdsprache befassen. Bisherige Untersuchungen sprachlicher Kompetenz, sowohl wissenschaftliche Studien (vgl. Erickson 2009; Skar 2013; Borger 2018) als auch von der Seite der schwedischen Schulbehörden (z. B. Skolinspektionen 2010; Skolverket[5] 2020b), haben hauptsächlich die Bewertung von Lernendenproduktionen in Schwedisch (L1) und Englisch (L2) untersucht. Nicht zuletzt die Bewertung schriftlicher Kompetenz ist mit Fragen der Validität bzw. der Interpretation von deren Ergebnissen konfrontiert. Darüber hinaus ist der Bereich Fremdsprachendidaktik hinsichtlich der zweiten Fremdsprache, d. h. in anderen Sprachen als Englisch, in der Forschung ein etwas vernachlässigtes Thema (vgl. Cabau-Lampa 2007; Bardel et al. 2016) und dies gilt im schwedischen Kontext insbesondere für das Fach Deutsch.

Bei der Auseinandersetzung mit Validität ist von Relevanz, inwiefern in den Urteilen der jeweiligen Bewertenden ähnliche oder unterschiedliche *Sprachkompetenzkonstrukte* reflektiert werden. Inwiefern die Bewertenden ihre Aufmerksamkeit auf dieselben oder ähnliche Aspekte richten oder inwiefern Aspekte bei einer Bewertung von einzelnen Bewertenden mehr Gewicht enthalten, ist infolge dessen bei einer validen Bewertung von großer Bedeutung. In einem schwedischen Schulkontext gibt es aber wenige wissenschaftliche Arbeiten zu Validitätsaspekten im Hinblick darauf, *wie* Bewertende das zu messende Konstrukt[6] konzeptualisieren und *welche* Aspekte sie bei einer Bewertung schriftlicher Kompetenz berücksichtigen (vgl. hierzu Borger 2018).

Die *Bewerterübereinstimmung* bei einer Bewertung, insbesondere zwischen Lehrkräften in den einzelnen Schulen und externen Bewertenden, hat die Aufmerksamkeit der schwedischen Schulbehörde (*Skolverket*) erregt und wird im schwedischen Schulkontext häufig diskutiert, insbesondere nach den in den Medien oft beachteten Zweitkorrekturen des schwedischen Schulinspektorats (z. B. Skolinspektionen 2010; 2018). Im Zentrum dieser Berichte steht die Bewerterübereinstimmung schwedischer Lehrkräfte, z. B. inwiefern Bewertende in ihren Bewertungen schriftlicher Leistungen zu möglichst ähnlichen Ergebnissen kommen. Eine hohe Reliabilität bedeutet jedoch nicht automatisch,

---

5 Schwedische Behörde für Schule und Erwachsenbildung (Nationale Agentur für Bildung).

6 Der Begriff *Konstrukt* bezieht sich auf „the concept or characteristic that a test is designed to measure“ (American Educational Research Association et al. 2014: 11). Ausgehend vom Testkonstrukt können Aufgaben und Bewertungskriterien unterschiedliche Fokusse hinsichtlich Verwendungskontexte und Aspekte der sprachlichen Kompetenz haben.

dass gleichzeitig eine hohe Validität vorliegt (vgl. Lumley 2002; Koretz 2008). Ein angemessenes Maß an Reliabilität bei einer Bewertung ist dahingegen aber eine Voraussetzung für die Validität (z. B. Erickson & Sylvén 2013).

Bei einer Untersuchung der Validität ist auch von Relevanz, in welchem Verhältnis die Testergebnisse einer Bewertung zu Einschätzungen sprachlicher Kompetenz, die in etwa die gleiche Kompetenz zeigen sollen, stehen. Als deutliche Inspirationsquelle für das schwedische System für Fremdsprachen dient der Referenzrahmen des Europarats. Die Fremdsprachenstufen, die schwedische Lernende am Gymnasium belegen, orientieren sich an bestimmten *Sprachniveaus des GER*. Es hat sich aber in den wenigen bisher durchgeführten empirischen Studien erwiesen, dass sich Sprachlernende der zweiten Fremdsprache im schwedischen Bildungssystem nicht immer tatsächlich auf dem zu erwartenden sprachlichen GER-Niveau befinden (vgl. European Commission 2012b; Granfeldt et al. 2019b; Aronsson 2020). Im Projekt *European Survey on Language Competences, ESLC,* wurden Fremdsprachenkenntnisse europäischer Jugendlicher in Englisch und der in jedem Land nach Englisch meistgewählten Fremdsprache (in Schweden: Spanisch) in 14 europäischen Ländern untersucht. Es hat sich dabei herausgestellt, dass schwedische Schülerinnen und Schüler am Ende der neunten Jahrgangsstufe in Englisch über sehr gute Sprachkenntnisse verfügen, dahingegen generell die Kompetenzerwartung des GER im Hinblick auf die zweite Fremdsprache Spanisch nicht erreichen (vgl. European Commission 2012b). Eine weitere Studie des sog. TAL-Projekts, einer größeren Forschungsstudie über Spracherwerb, Sprachunterricht und mündliche Sprachleistungen der drei meistgewählten Sprachen Deutsch, Französisch und Spanisch im Fach *Moderna språk*, hat darauf verwiesen, dass schwedische Schülerinnen und Schüler am Ende der neunten Jahrgangsstufe im Hinblick auf die mündliche Sprachfertigkeit nicht das erwartete GER-Niveau erreichen, was dementsprechend auch für die Fremdsprachenkenntnisse in Deutsch gilt (vgl. Granfeldt et al. 2019b).

Auch wenn nicht alle Teilkompetenzen in den bisherigen Studien untersucht wurden, scheinen insbesondere die Fremdsprachenkenntnisse im Hinblick auf die schriftliche Kompetenz unter dem zu erwartenden GER-Niveau am Ende der schedischen Grundschule[7] zu liegen (vgl. European Commission 2012b;

7 Die Struktur des schwedischen Schulwesens ist anders organisiert wie z. B. in Deutschland. Die schwedische einheitliche Grundschule (*grundskola*) kann als Ganztagsschule beschrieben werden und die Schulpflicht umfasst heute 10 Jahre. Nach der Grundschule können schwedische Schülerinnen und Schüler freiwillig das dreijährige Gymnasium besuchen, wo sie die Hochschulreife erhalten können. Fast

Aronsson 2020). Davon abgesehen haben zudem nur einzelne Studien die Kompetenzniveaus für Lernproduktionen mit höheren Noten als der ausreichenden Note E untersucht und dabei geprüft, in welchem Verhältnis diese höher benoteten Leistungen zum GER stehen (vgl. Erickson 2019). Bisherige Studien in der zweiten Fremdsprache haben aber hauptsächlich die Sprachkompetenz von Lernenden am Ende der Grundschule untersucht, weshalb empirische Daten über die Sprachkompetenzen am Gymnasium nicht vorliegen. Dies insgesamt könnte ein Anlass sein, gerade die schriftliche Sprachkompetenz von Deutschlernenden am Gymnasium zu untersuchen.

Hierbei lässt sich die Frage stellen: In welcher Beziehung stehen Bewertungen schwedischer Bewertender auf unterschiedlichen Fremdsprachenstufen des schwedischen Gymnasiums zu Bewertungen hinsichtlich eines bestimmten GER-Niveaus? Diese Frage ist insbesondere deswegen relevant, weil die Kompetenzniveaus von schwedischen Sprachlernenden mit den Niveaus von Lernenden in anderen Ländern verglichen werden und ein erreichtes GER-Niveau auch für die Aufnahme eines Studiums oder für Karrierewege im Ausland wichtig sein kann. Als Basis für die Einschätzung allgemeiner Kompetenzniveaus eines Lernenden werden häufig eher die produktiven als die rezeptiven Kompetenzen bevorzugt. Die direkt getesteten produktiven Kompetenzen werden als „a more relevant, practical and meaningful target for aligning judgements of level across classroom and large-scale assessments“ (Jones & Saville 2016: 74) wahrgenommen. Die Tatsache, dass Sprachlernende erfahrungsgemäß sowie laut bisherigen Studien (vgl. Lenz & Studer 2008; European Commission 2012b) zudem in produktiven Kompetenzen ein niedrigeres Sprachniveau als in rezeptiven aufweisen und die schriftliche Kompetenz bisher nur in wenigen Studien untersucht wurde (z. B. European Commission 2012b), ist ein Grund für den Fokus auf die schriftliche Kompetenz in der vorliegenden Arbeit.

Es ist bei der gewählten Fokussierung von großem Gewicht, die Bewertung schriftlicher Kompetenz umfassend und aus unterschiedlichen Perspektiven zu untersuchen. Dass wir bislang wenig über eine Bewertung schriftlicher Kompetenz in der zweiten Fremdsprache und darüber, wie wir diese Testergebnisse interpretieren können, wissen, kann daher als ein Problem aufgefasst werden. Vor diesem Hintergrund erscheint es von umso größerer Relevanz, auch Validitätsaspekte hinsichtlich der Konstruktkonzeptualisierung und der Bewerterübereinstimmung von Bewertenden bei der Bewertung schriftlicher

alle Schülerinnen und Schüler in Schweden wechseln von der Grundschule auf das Gymnasium.

Kompetenz in einer zweiten Fremdsprache ausgehend von einem schwedischen Schulkontext zu untersuchen.

Der GER als Bezugspunkt

Die Bewertung sprachlicher Lernergebnisse und Kompetenzen, sowohl im schulischen als auch außerschulischen Kontext, orientiert sich in immer höherem Grad an externen Referenzsystemen. Das in Europa meistverbreitete Referenzsystem für das Erlernen und die Bewertung von Fremdsprachenkenntnissen ist der bereits erwähnte gemeinsame Referenzrahmen für Sprachen (vgl. Europarat 2001), an dem sich Prüfungen und Bildungsstandards orientieren können. Der GER verfolgt das Ziel, fremdsprachliche Kompetenzen zu beschreiben, und mit sechs festgelegten Niveaus[8] bietet der GER ein Referenzsystem für die Einschätzung von sprachlichen Kompetenzen. Diese Kompetenzstandards oder Niveaubeschreibungen sollten vor allem die Vergleichbarkeit von Sprachprüfungen und Bildungssystemen in Europa erleichtern und damit zu einer höheren Mobilität von Menschen und einer verstärkten interkulturellen Kommunikation in Europa führen. An diesem Punkt hatten die Referenzniveaus des GER zudem eine bedeutende Wirkung auf das Verständnis davon, was Lernende auf unterschiedlichen Niveaus in ihren L2 ausdrücken können.

Seit der Herausgabe des Referenzrahmens hat das Dokument auf Spracherlernen und Sprachpolitik in Europa einen sehr großen Einfluss ausgeübt. Auch wenn politische Entscheidungen in einzelnen Ländern dazu geführt haben, dass der GER in unterschiedlichem Grad implementiert ist, gilt er häufig als „the most significant recent event on the language education scene in Europe" (Alderson 2005: 257). Darüber hinaus wird der Referenzrahmen zunehmend als Basis für das Erstellen von Lehrwerken, Lehrplänen und Bildungsstandards verwendet. Dabei wurde beabsichtigt, dass der GER ein Bezugspunkt für alle beteiligten Partner bei der Beschreibung fremdsprachlicher Kompetenz sein sollte (Europarat 2001: 32). Der Referenzrahmen wird somit als gemeinsames Referenzsystem nicht nur für Testinstitute, sondern auch für die im Fremdsprachenunterricht vermittelte Kompetenz in den Bildungssystemen der jeweiligen europäischen Länder verwendet. Hierbei wird deutlich, dass die Referenzniveaus des GER eine Bedeutung als externer Bezugspunkt gehabt hat: „CEFR

8 Um Missverständnisse zu vermeiden, bezieht sich das Wort *Niveau* in der vorliegenden Arbeit hauptsächlich auf die Referenzniveaus des GER und die Bezeichnungen *Stufe* bzw. *Fremdsprachenstufe* beziehen sich primär auf die verschiedenen Stufen des schwedischen Bildungssystems.

has had an enormous effect on how L2 teaching and assessment relate to and can be aligned with a set of external standards“ (Purpura 2016: 202). Der Bedarf an gemeinsamen, externen Richtlinien, die zugleich möglichst transparent und objektiv verfolgt werden können, scheint für Einschätzungen von Sprachkompetenzen über die Landesgrenzen hinaus größer als jemals zuvor zu sein. Darüber hinaus standen in den letzten Jahren zunehmend Evaluierungen von generellen Schülerleistungen, erwarteten Leistungsniveaus sowie Schulen und Schulsystemen im Zentrum nationaler und internationaler bildungspolitischer Diskussionen.[9] Vor diesem Hintergrund werden die Referenzniveaus des GER zunehmend als eine Grundlage für Bestimmungen sprachlicher Niveaus von Lernenden verstanden.

Trotz der weiten Verwendung des Referenzrahmens, vor allem in Bezug auf die Referensniveaus, kann nicht außer Acht gelassen werden, dass das Dokument viel Kritik enthalten hat. Hierzu gehören Kritikpunkte hinsichtlich mangelnder Beschreibungen der Leistungsdeskriptoren (z. B. Harsch 2006) und einer normierenden Verwendung und unreflektierten Implemetierung des GER in nationalen Bildungssystemen (vgl. Quetz & Vogt 2009; Erickson 2011a). Auch wenn gegenüber dem GER Kritik geäußert wurde, ist durch den Referenzrahmen ein Bezugspunkt geschaffen worden, an dem sich Sprachprüfungen und Bildungsstandards orientiert haben. Vor diesem Hintergrund ist es erstaunlich, dass es vergleichsweise wenig empirische Forschung gibt, die das Verhältnis zwischen Lehrwerken und Lehrplänen auf einem gewissen Niveau und den entsprechenden Leistungsniveaus des GER untersucht hat. Dieser Mangel an Qualitätssicherung ist bereits vor vielen Jahren in der Forschung festgestellt worden:

> examination providers, textbook publishers, and curriculum developers make claims about the relationship between their products and the CEFR. […] The problem is that there is little empirical evidence to back up these claims (Alderson 2007: 661)

---

9 Der Forschungsbereich standardisierter Leistungsuntersuchungen gewinnt international an Bedeutung. Sprachleistungsstudien, vor allem hinsichtlich der zweiten Fremdsprache, kommen jedoch auf internationaler Ebene seltener vor als Leistungsstudien mit dem Fokus auf Lesekompetenzen, naturwissenschaftliche Kompetenzen oder mathematische Kompetenzen. Eine Ausnahme in Europa ist die Studie *European Survey on Language Competences*, (ESLC) eine erste europäische Erhebung von Sprachkompetenzen, die von der Europäischen Kommission (European Commission 2012b) durchgeführt wurde. Als Bezugspunkt sprachlicher Niveaus wurden in dieser Studie die sechs Referenzniveaus des GER verwendet.

Diese Aufforderung wurde aber im Bereich des Fremdsprachentestens aufgenommen. Im letzten Jahrzehnt hat die Zuordnung von Testergebnissen standardisierter Prüfungen zu den Niveaustufen des GER zugenommen, häufig gemäß den Richtlinien des von Europarat herausgegebenen Dokuments, des *Manuals* (Council of Europe 2009).

Während die Anbindung von Sprachtests an den GER die Forschung in den letzten Jahren dominiert hat (Papageorgiou, 2016: 329), ist aber der Bezug des GER zu europäischen Bildungssystemen bisher vernachlässigt worden. Darunter fällt auch die Zuordnung der Fremdsprachenstufen des schwedischen Bildungssystems zu den Referenzniveaus des GER. In einem Bericht der Europäischen Kommission aus dem Jahr 2013 wird insbesondere auf den Mangel an empirischen Belegen für einen Zusammenhang zwischen den Kompetenzstufen des GER und den Lernergebnissen verwiesen:

> Dennoch gibt es einige Bedenken, was die Umsetzung des GER in Schweden anbelangt: Der Zusammenhang zwischen den für das Bildungswesen geltenden Rechtsdokumenten und dem GER wird von keiner wissenschaftlichen Studie empirisch abgesichert (Broek & van den Ende 2013: 71)

Die mangelnden empirischen Belege hinsichtlich Lernergebnisse sind jedoch nicht das einzige Bedenken. Schweden hat eine deutlich vorsichtige Haltung dem GER gegenüber eingenommen, wie auch Länder wie Norwegen und Dänemark (vgl. Erickson & Pakula 2017), was bedeutet, dass die mögliche Bedeutung des GER für das Lernen von Fremdsprachen und dessen explizite Verwendung im schulischen Fremdsprachenunterricht weniger ausgesprochen ist. Diese etwas vorsichtige Haltung kann jedoch sowohl Vorteile als auch Nachteile mit sich bringen. Obwohl das zu Beginn des 21. Jahrhunderts eingeführte siebenstufige Modell des schwedischen Systems für Englisch und die modernen Sprachen,[10]

---

10 Das Fach Englisch und die Fächer der *Moderna språk* haben im schwedischen System dieselbe Struktur mit sieben Stufen, die sich an den Referenzniveaus des GER orientieren. Jede Stufe identifiziert und beschreibt die sprachliche Kompetenz eines Lernenden unabhängig von der gelernten Sprache, z. B. Englisch, Deutsch, Französisch oder Spanisch. Die Fächer der *Moderna språk* hat demzufolge im schwedischen Bildungssystem eine Progression von sieben aufeinander aufbauenden Stufen, von 1 bis 7. In den nationalen Rahmenplänen für *Moderna språk* werden keine einzelnen Sprachen angegeben. Die Stufen werden folglich als *Moderna språk 1* („Moderne Sprache 1"), *Moderna språk 2* („Moderne Sprache 2"), usw. bezeichnet. Im Folgenden werden die schwedischen Bezeichnungen *Tyska 1*, *Tyska 2*, *Tyska 3*, *Tyska 4*, *Tyska 5*, *Tyska 6*, *Tyska 7* verwendet, wenn auf die Fremdsprachenstufen im Fach Deutsch (*Tyska*) im schwedischen Bildungssystem verwiesen wird, um die jeweilige Stufe

das von bisherigen Publikationen des Europarates beeinflusst war, seit über 20 Jahren vorliegt, gibt es folglich bisher kaum Studien darüber, in welchem Verhältnis Sprachkompetenzen in den jeweiligen Fremdsprachenstufen des schwedischen System zu einem bestimmten Referenzniveau des GER stehen. Eine empirische Untersuchung mittels Bewertungen erfahrener Lehrkräfte mit für den Zweck geeigneten Testdaten wäre daher bedeutsam, insbesondere im Hinblick auf die weltweit verstärkte Bedeutung des GER als Bezugssystem. In der vorliegenden Arbeit wird versucht, diese Lücke in der Forschung zu schließen, indem ausgewertet wird, inwiefern Schülerinnen und Schüler aus verschiedenen Fremdsprachenstufen am schwedischen Gymnasium, die Anforderungen im schriftlichen Teil eines Tests auf einem bestimmten GER-Niveau erfüllen, und ihre Resultate mit einer schwedischen Bewertung nach den nationalen Standards verglichen werden.

1.1 Zielsetzung und Fragestellungen

Die vorliegende Arbeit hat zum Ziel, Validitätsaspekte bei der Bewertung schriftlicher Sprachkompetenz im Fach Deutsch zu untersuchen, wobei der schwedische Schulkontext als Ausgangspunkt dient. Die schwedische Sichtweise auf Unterricht und Bewertung in einer Fremdsprache, in den Bildungsstandards festgelegt, ist durch den handlungsorientierten Ansatz des GER geprägt. Somit liegt im schwedischen System die Betonung eher auf einer Leistungsbewertung, die eine kommunikative Sprachfähigkeit im weiteren Sinne prüfen soll und bei der sich die Lehrkräfte in ihren Entscheidungen nach kriterienorientierten Lernzielen richten, als auf einer Leistungsmessung mit detailliert überprüfbaren Direktiven und Richtlinien. Da die Lehrkräfte im schwedischen System im internationalen Vergleich eine verhältnismäßig hohe Verantwortung für die Bewertung haben und da Abschlussprüfungen mit externen Bewertenden im Gegensatz zu vielen anderen Ländern in der Regel nicht vorkommen, ist eine Untersuchung der Validität bei der Bewertung fremdsprachlicher Kompetenz von großer Relevanz. Die Studie fokussiert aus diesem Grund auf Validitätsaspekte bei der Bewertung *nach* dem Testereignis, *a posteriori* (vgl. Weir 2005). Bei Untersuchungen der Validität soll nicht der Test im Vordergrund stehen, sondern die Interpretation und die Verwendung der Testergebnisse (vgl. Messick 1989b). Gemäß Kane ist von Bedeutung, dass relevante Aspekte der Validität

---

klarzustellen und dabei auch zu verdeutlichen, dass das Fach Deutsch in der vorliegenden Arbeit im Fokus steht.

untersucht werden und dass in vorgegebenen Schritten eines Validitätsmodells unterschiedliche Nachweise für die Interpretation und Verwendung der Testergebnisse eingeholt werden können (vgl. Kane 2013).

Diese Arbeit stellt somit eine Studie zur Validität bei einer Bewertung fremdsprachlicher Schreibkompetenz von schwedischen Schülerinnen und Schülern am Gymnasium im Fach *Tyska* (Deutsch) dar. Untersucht wird die Bewertung fremdsprachlicher Textproduktionen auf den Fremdsprachenstufen *Tyska 3*, *Tyska 4* und *Tyska 5*. Die für diese Studie fokussierte Bewertung von Sprachkompetenzen wird im Hinblick auf verschiedene für einen schwedischen Schulkontext relevante Validitätsaspekte untersucht: a) die Konstruktkonzeptualisierung der Bewertenden, b) die Übereinstimmung schwedischer Bewertender und c) die Beziehung schwedischer Bewertungen zu einem bestimmten Sprachniveau des GER bei der Bewertung schriftlicher Kompetenz.

Die Fragestellungen zielen dabei auf verschiedene Schritte einer Bewertung schriftlicher Schülerleistungen ab. Die Untersuchung bezieht sich auf folgende drei Hauptfragen, die im Folgenden präzisiert werden:

1. Welche Aspekte auf der Ebene der Texte sind in den jeweiligen Bewerterurteilen besonders relevant für die Beurteilung und wie unterscheiden sich die Urteile zwischen einzelnen Bewertenden und Bewertergruppen bezogen auf: a) die eigene Lehrkraft, b) die externen schwedischen Bewertenden sowie c) die GER-Bewertenden?
2. Wie unterscheiden sich Bewertungen bezüglich der Bewerterübereinstimmung unter den schwedischen Bewertenden?
3. In welcher Beziehung stehen Bewertungen von Textproduktionen schwedischer Schülerinnen und Schüler auf den Fremdsprachenstufen *Tyska* 3, *Tyska* 4 und *Tyska* 5 des schwedischen Bildungssystems zu Bewertungen der schriftlichen Sprachkompetenz auf einem erfüllten B1-Niveau des GER?

Die Fragestellungen beziehen sich auf das empirische Material der Untersuchung. Hierbei soll u. a. herausgearbeitet werden, welche Aspekte Bewertende bei ihrer Bewertung bezüglich der schriftlichen Kompetenz für besonders wichtig erachten, inwieweit eine zuverlässige Bewertung der Textproduktionen von schwedischen Bewertenden gewährleistet werden kann und inwiefern die Testergebnisse schwedischer Bewertungen als ein Indikator für die schriftliche Kompetenz hinsichtlich eines GER-Niveaus B1 betrachtet werden können. Die vorliegende Untersuchung soll somit zum besseren Verständnis des Bewertungsprozesses und der Verwendung und Interpretationen der daraus abgeleiteten Testergebnisse schriftlicher Leistungen in einer Fremdsprache führen. Zum einen können die Ergebnisse der Studie wichtige Informationen im

Hinblick darauf geben, welche Aspekte bei der Bewertung schriftlicher Kompetenz Berücksichtigung finden sowie Tendenzen hinsichtlich der Bewerterübereinstimmung schwedischer Bewertender offenbaren. Zum anderen lässt sich mit Bezug auf ein bestimmtes Referenzniveau des GER untersuchen, in welchem Verhältnis schwedische Bewertungen zu einem externen Referenzniveau des GER im Hinblick auf die schriftliche Kompetenz stehen. Diese Untersuchung kann somit in gewissem Ausmaß einen gewissen Beitrag zur empirischen Anbindung der Fremdsprachenstufen in Schweden an ein bestimmtes GER-Niveau leisten.

Es ist an dieser Stelle aber wichtig zu erwähnen, dass die vorliegende Studie aufgrund der relativ kleinen Untersuchungsunterlage nur einen Hinweis auf die Berücksichtigung relevanter Aspekte bei der Bewertung, auf die Reliabilität jener Bewertung sowie auf die Beziehung zwischen den jeweiligen Fremdsprachenstufen und einem B1-Niveau des GER geben kann. Die vorliegende Arbeit ermöglicht es aber, einen Blick auf relevante Validitätsaspekte hinsichtlich der Bewertung schriftlicher Kompetenz von Schülerinnen und Schülern im Fach *Tyska* im schwedischen System zu werfen. Darüber hinaus kann diese empirische Untersuchung als ein erster Schritt eines Validierungsprozesses für die Zuordnung fremdsprachlicher Leistungen von Schülerinnen und Schülern in Deutsch am schwedischen Gymnasium verstanden werden. Die Arbeit wendet sich hierbei an ein breites Publikum, u. a., Lernende, Lehrkräfte, Lehramtstudierende, Schulleitende, Forschende sowie andere Akteure im Bildungsbereich, die an fachdidaktischen Fragen hinsichtlich einer Bewertung interessiert sind. Gleichzeitig kann die Untersuchung durch den systematischen und theoretischen Validierungsansatz zum internationalen wissenschaftlichen Diskurs beitragen.

1.2 Aufbau der Arbeit

Auf die Einleitung, in der Problematik, Zielsetzung und Fragestellungen erklärt werden (Kap. 1), folgen eine Kontextualisierung und eine Darstellung des Hintergrunds zum Fremdsprachenunterricht in Schweden. Hierbei wird auf die Entstehung und den Einfluss des GER sowie die schwedischen Bildungsstandards für Fremdsprachen und deren Anbindung an den Referenzrahmen eingegangen (Kap. 2). Danach wird der konzeptionelle Rahmen behandelt, innerhalb dessen die Fragestellungen der vorliegenden Arbeit verfolgt werden (Kap. 3), sowie der Stand der Forschung im Hinblick auf Relevanz für die Arbeit erläutert (Kap. 4). Des Weiteren werden das Forschungsdesign und die Methodenwahl dargestellt. Hierbei werden auch die Datenerhebung des empirischen

Materials sowie die angewandten Methoden und Testinstrumente präsentiert (Kap. 5).

In weiteren Kapiteln folgt die Darstellung der empirischen Untersuchung. Hierbei werden die Ergebnisse zu Schwerpunktsetzungen bei der Bewertung schriftlicher Kompetenz in einer Fremdsprache (Kap. 6) und zur Bewerterübereinstimmung der schwedischen Bewertenden (Kap. 7) dargelegt sowie die Beziehung der Bewertungen schriftlicher Sprachkompetenzen schwedischer Deutschlernenden zum angestrebten GER-Niveau B1 (Kap. 8). In der sich anschließenden Diskussion werden von einem Validierungprozess in mehreren Schritten ausgegangen und verschiedene Validitätsaspekte bei der Bewertung fremdsprachlicher Kompetenzen erörtert (Kap. 9). Abschließend erfolgt eine kurze Zusammenfassung der wichtigsten empirischen Befunde, eine Auslotung der Grenzen der Studie sowie ein Ausblick auf weitere Forschungsperspektiven und didaktische Implikationen sowohl für das Lernen und Lehren einer fremden Sprache als auch für die Bewertung fremdsprachlicher Kompetenz (Kap. 10).

2. Kontextueller Hintergrund

Um den schwedischen Deutschunterricht in einen Kontext einzuordnen, soll zunächst ein historischer Abriss erfolgen, an den anschließend der aktuelle Stand des Fremdsprachenlernens in Schweden, das schwedische System für den Fremdsprachenunterricht, der Einfluss des Referenzrahmens und dessen Bezug zu schwedischen Bildungsstandards erörtert werden. Der Fokus hierbei liegt hauptsächlich auf dem gegenwärtigen System für das Erlernen einer zweiten Fremdsprache[11] in der schwedischen Schule. Da die vorliegende Arbeit einen Beitrag zum Verständnis der Zuordnung von Sprachkompetenzen von Lernenden zu den Referenzniveaus des GER leisten möchte, wird in diesem Kapitel auch der Referenzrahmen kurz beschrieben und auf dessen Bedeutung als Referenzpunkt eingegangen.

Im ersten Abschnitt wird zunächst ein kurzer Überblick über die Entwicklung des Schulfachs Deutsch im schwedischen Bildungssystem gegeben (Kap. 2.1). Danach werden Organisation und Aufbau des Fremdsprachenunterrichts, aktuelle Informationen über die Verteilung der Lernenden im Hinblick auf die zweite Fremdsprache (insbesondere für das Fach Deutsch) sowie die Voraussetzungen einer Bewertung im Fach wiedergeben. Im Mittelpunkt der vorliegenden Arbeit steht die Bewertung fremdsprachlicher Schreibkompetenz im schwedischen Schulkontext. Welche Kompetenzen müssen schwedische Schülerinnen und Schüler im Hinblick auf die Anforderungen in den Bildungsstandards hinsichtlich Schreibkompetenz erfüllt haben und wie werden diese Kompetenzen beschrieben? Im Weiteren folgt zunächst eine kurze Beschreibung der schwedischen Lehrpläne für *Moderna språk* sowie deren zentralen Inhalten und Anforderungen hinsichtlich der schriftlichen Kompetenz, insbesondere im Hinblick auf die in der vorliegenden Studie untersuchten Fremdsprachenstufen im schwedischen System. Dazu wird auf das Thema Bewertung und fakultative Tests in der zweiten Fremdsprache sowie auf gegenwärtige

11 Die Bezeichnungen „zweite Fremdsprache" bzw. „moderne Sprache" (*Moderna språk*) beziehen sich in der vorliegenden Arbeit auf die weiteren gegenwärtigen Fremdsprachen, die nach der ersten Fremdsprache Englisch in der schwedischen Schule gelernt werden. In der Regel ist die nach Englisch gewählte Fremdsprache auch die zweite Fremdsprache und daher werden diese Bezeichnungen synonym gebraucht. Diese werden hier verwendet, damit zwischen der erster Fremdsprache Englisch und den weiteren Fremdsprachen in der schwedischen Schule unterschieden werden kann.

bildungs- und sprachpolitische Maßnahmen und Diskussionen eingegangen, die für den Fremdsprachenunterricht in Schweden von Bedeutung sind (Kap. 2.2).

Darauffolgend werden kurzgefasst der Entstehungsprozess, die Grundlagen und die Auswirkung des *Gemeinsamen europäischen Referenzrahmen für Sprachen* (Europarat 2001) für den Fremdsprachenerwerb, den Fremdsprachenunterricht und für das Beurteilen von Fremdsprachen in Europa erörtert. Darüber hinaus wird der GER als Bezugssystem betrachtet, u. a. im Hinblick auf die Validierung der Anbindung von Bildungsstandards und internationalen Sprachtests an den Referenzrahmen (Kap. 2.3). Abschließend wird auf die Bedeutsamkeit und Umsetzung des GER in Schweden Bezug genommen. Dabei wird auf die schwedischen Bildungsstandards für Fremdsprachen sowie deren Bezug zum Referenzrahmen eingegangen. In Verbindung damit werden die Fremdsprachenstufen des schwedischen Systems mit den Referenzniveaus des GER in Beziehung gestellt, wie es im sog. Kommentarmaterial zum Lehrplan für das Fach *Moderna språk* (Skolverket 2011b[12]) dargestellt ist (Kap. 2.4).

2.1 Deutsch als Schulfach in Schweden

Fremdsprachenunterricht und insbesondere das Erlernen von Deutsch (*Tyska*) hat im schwedischen Kontext eine lange Tradition, auch wenn sich anfangs nur eine winzige Minderheit aller Kinder im Schulalter Sprachstudien widmeten und diese oft als Privatunterricht durchgeführt wurden.[13] Im 19. Jahrhundert erfolgte der Durchbruch der modernen Sprachen, die damit die vorherrschende Dominanz der klassischen Sprachen im schwedischen System abgelöst haben. Fortan konnten Deutsch und Französisch – und in gewissem Ausmaß auch Englisch – am Gymnasium gewählt werden. Ab Mitte des 19. Jahrhunderts wurden die drei modernen Sprachen Deutsch, Französisch und Englisch somit

12 Dieses Verhältnis zwischen den Fremdsprachenstufen und den GER-Niveaus ist auch in der neuen Fassung des Kommentarmaterials für die schwedischen Lehrpläne im Fach *Moderna språk* und Englisch dargestellt (vgl. Skolverket 2021e). Hier und im Folgenden wird die zum Zeitpunkt der Datenerhebung aktuelle Version des Kommentarmaterials aus dem Jahr 2011 verwendet.

13 Einen Überblick über die historische Entwicklung des schulischen Fremdsprachenunterrichts in Schweden gibt beispielsweise Cabau-Lampa (2005). Des Weiteren beschreibt Bernhardsson (2016) das wechselnde Verhältnis zwischen Privatunterricht und Schulunterricht für das Erlernen einer modernen Fremdsprache im 19. Jahrhundert.

zum festen Bestandteil der schwedischen Schule und Schülerinnen und Schüler, die diese Sprachen erlernen wollten, wurden nicht mehr nur auf Privatunterricht verwiesen. Deutsch wurde bald zur ersten erlernten Fremdsprache und sollte für ein ganzes Jahrhundert die Rolle als wichtigste Fremdsprache in schwedischen Schulen halten, gefolgt von Französisch und auf dem dritten Platz Englisch (Malmberg 1986).

In der Lehre der Fremdsprachen Deutsch und Französisch wurde anfangs die Methodik aus der Lehre von klassischen Sprachen wie Latein oder Griechisch übernommen. Es ging hauptsächlich um das formelle Üben von Grammatik und Übersetzungen mit grammatischer Analyse, gemäß der sog. Grammatik-Übersetzungsmethode. Allerdings mehrten sich gegen Ende des 19. Jahrhunderts kritische Stimmen: Der Fokus sollte auf die lebendige gesprochene Sprache anstatt der geschriebenen gelegt werden. Verfechter solcher Entwicklungen in der Fremdsprachenlehre meinten auch, dass das Üben von Übersetzungen durch freie mündliche und schriftliche Produktion ersetzt werden sollte. Die Grammatik-Übersetzungsmethode war dennoch bis in das 20. Jahrhundert hinein in den schwedischen höheren Schulen zu sehen (vgl. Malmberg 1986).

Eine Fremdsprache zu lernen war jedoch immer noch nicht obligatorisch und wurde nicht von allen Schulformen angeboten. Erst die Schulkommission aus dem Jahr 1946 griff dies auf und schärfte das Bewusstsein für die Bedeutung von Fremdsprachenkenntnissen weiter. Eine gemeinsame Schule für alle wurde hier als ein Teil eines Demokratisierungsprozesses angesehen. Laut der Schulkommission gehörte das Erlernen von Fremdsprachen zu diesem Demokratisierungsprozess:

> Ein aus staatsbürgerlicher Sichtweise spürbares Defizit in der bisherigen Aufstellung der Schulfächer der obligatorischen Schule ist die Abwesenheit von Fremdsprachenunterricht. Fremdsprachenkenntnisse wurden bisher nur wenigen vorbehalten, den sog. Gebildeten. Wenn eine Fremdsprache – und die Wahl wird dann mit gutem Grund auf Englisch fallen – als Pflichtfach in der Pflichtschule eingeführt werden sollte, sollte dadurch eine alte Bildungskluft zumindest erträglich überbrückt werden. Ein Fenster zur Welt würde für die breite Masse der Bürger geöffnet werden. Zunehmend setzen sich auch Kenntnisse von zumindest einer Fremdsprache in sowohl Berufs- als auch Organisationsleben durch.[14] (SOU 1948:27, S. 7, *eigene Übersetzung, M.H.R*)

---

14 Im Original: „En ur medborgerlig synpunkt kännbar brist i den obligatoriska skolans hittillsvarande ämnesuppsättning är frånvaron av undervisning i främmande språk. Kunskaper i främmande språk har hittills varit förbehållna ett litet fåtal, de s. k. bildade. Om ett främmande språk – och valet kommer då rimligen att falla på engelskan – införes som obligatoriskt ämne i skolpliktstidens skola, skulle därigenom en

Hier wird deutlich, dass das Erlernen einer Fremdsprache fortan nicht für eine gebildete Minderheit reserviert sein sollte; dem zunehmenden Bedarf an Sprachkenntnissen im Berufs- und Organisationsleben sollte somit entgegenkommen werden. Als erste obligatorische Fremdsprache sollte nach Ansicht der Kommission Englisch angeboten werden. Die Schulkommission wollte auch die Zielsetzung für die Methodik im Fremdsprachenunterricht ändern. Hauptziel für den Unterricht sollte sein, Texte in der Fremdsprache zu lesen und zu verstehen, gefolgt von Hörverstehen und aktiver Teilnahme in Gesprächen mit Muttersprachlern. Die schriftliche Kompetenz sollte im Unterricht eher eine untergeordnete Rolle einnehmen und insbesondere nicht länger in Form von Übersetzungsübungen geübt werden (SOU 1948:27, S. 29).

Ein obligatorischer Unterricht in Englisch wurde in Schweden auch nach der Empfehlung der Kommission nicht gleich eingeführt. Im Jahr 1946 wurde aber Englisch die erste Fremdsprache ab Jahrgang 5; als weitere Fremdsprachen konnten Deutsch ab Jahrgang 7[15] und Französisch ab Jahrgang 9 gewählt werden. Erst seit 1962 ist Englisch in der schwedischen Grundschule Pflichtfach. Im selben Jahr wurde Deutsch mit Französisch gleichgestellt, indem zwischen diesen beiden Sprachen als zusätzlicher Option in der Grundschule gewählt werden konnte. Innerhalb von studienvorbereitenden Ausrichtungen war allerdings das Erlernen einer zweiten Fremdsprache obligatorisch. Als zweite Fremdsprache hat Deutsch danach bezüglich der Beliebtheit der beiden Sprachen für eine lange Zeit über Französisch dominiert (Cabau-Lampa 2005). Mit der Einführung von Englisch als erste obligatorische Sprache für alle Kinder der schwedischen Schule wurde 1962 eine neue Pädagogik im Fremdsprachenunterricht verlangt. Bereits die Schulkommission aus dem Jahr 1946 hat aber die Frage der Methode im Fremdsprachenunterricht aufgegriffen:

> Der grammatisch ausgerichtete Sprachunterricht, der unsere Schule so stark dominiert hat, litt unter dem pädagogischen Irrtum, übermäßige Anforderungen an die intellektuellen Voraussetzungen der Anfänger zu stellen. Indem Deutsch durch Englisch als erste Fremdsprache in unseren Schulen ersetzt wird, haben die pädagogischen

---

gammal bildningsklyfta åtminstone hjälpligt överbryggas. Ett fönster ut mot världen skulle öppnas för den breda massan av medborgare. I allt högre grad gör sig också behovet av kunskaper i åtminstone ett främmande språk gällande både i yrkes- och organisationslivet.“

15 Es wird dabei aber von der Schulkommission aus dem Jahr 1946 angenommen, dass eine große Anzahl von Schülerinnen und Schülern Deutsch als Fremdsprache in der Oberstufe lernen werden, da verschiedene Arten von Weiterbildungen Fremdsprachenkenntnisse in Deutsch erfordern.

> Möglichkeiten, den Sprachunterricht so zu gestalten, dass er für Kinder verständlicher wird, wesentlich zugenommen. Der Schwerpunkt kann nun auf das Lesen von einfacheren Texten, imitative Sprachübungen, Sprechübungen und andere […] konkretere, lebendigere und für Kinder interessantere Arbeitsmethoden gelegt werden.[16] (SOU 1948:27, S. 66, *eigene Übersetzung, M.H.R*)

Hier wird deutlich, dass die Schulkommission auf eine Veränderung bezüglich der Methodik im Fremdsprachenunterricht hoffte. Mit dem Einführen von Englisch verband sich offenbar eine Hoffnung, dass sich Unterrichtsmethoden mit eher formellen Wurzeln aus dem Unterricht klassischer Sprachen in einen Unterricht mit Schwerpunkt auf das Lesen geeigneterer Texte und die gesprochene Sprache umwandeln würden.

Die Sichtweise auf den Fremdsprachunterricht in Schweden hat sich in den 70er Jahren verändert. Von einem bisherigen Schwerpunkt auf sprachliche Form wechselt der Fokus jetzt auf die Funktion der Sprache (Erickson & Sylvén 2013), auch wenn gemäß Malmberg (1986) bereits vorher ein wechselhaftes Verhältnis zwischen den beiden Polen Form und Funktion zu beobachten war.[17] Die Mindestanforderungen der Bildungsstandards sollten sich parallel zu internationalen Entwicklungen nicht mehr an grammatischen Strukturen,

---

16 Im Original: „Den grammatiskt inriktade språkundervisningen, som så starkt dominerat i vår skola, led av det pedagogiska felet att ställa alltför stora krav på nybörjarnas allmänna intellektuella förutsättningar. I och med att tyska ersatts av engelska som första främmande språk i våra skolor, har de pedagogiska möjligheterna att lägga språkundervisningen på ett för barn mera fattbart sätt väsentligt ökat. Tonvikten kan nu läggas på läsning av enkel text, på imitativa språkövningar, talövningar och andra […] mera konkreta, levande och för barn mer intressanta arbetssätt."

17 In der Forschung und im Unterrichtsbereich hinsichtlich Fremdsprachen wird häufig zwischen deklarativen und prozeduralen Sprachkenntnissen unterschieden. Unter deklarativen Sprachkenntnissen werden in diesem Zusammenhang Erklärungen sprachlicher Phänomene und Kenntnisse sprachlicher Regeln verstanden. Prozedurale Sprachkenntnisse kennzeichnen die Sprachfertigkeiten, die ein Lernender in einer kommunikativen Situation verwenden kann (vgl. Tornberg 2015). Tornberg unterscheidet dabei zwischen einer produktausgerichteten Sichtweise der Grammatik mit Fokus auf sprachliche Form und einer prozessausgerichteten Sichtweise der Grammatik mit Fokus auf Sprachverwendung. Diese beiden Typologien sind nach Tornberg (ibid.) im Fremdsprachenunterricht notwendig; im Unterricht formreicher Schulsprachen, wie Deutsch und Französisch, sollte jedoch aus Tradition ein Schwerpunkt auf Grammatik als Produkt gelegt werden, im Vergleich zum Unterricht in Englisch, wo eher die Prozessperspektive überwiegend ist.

sondern eher an kommunikativ ausgerichteten Kriterien orientieren.[18] Der handlungsorientierte Ansatz und der Fokus auf kommunikative Kompetenzen im Fremdsprachenunterricht kommt bereits im Lehrplan aus dem Jahr 1980 zum Vorschein (vgl. Erickson 2019) und hat seitdem einen deutlichen Einfluss auf das Lehren, Lernen und Bewerten von Fremdsprachen im schwedischen Schulkontext gehabt.

In den letzten Jahrzehnten sind eine Reihe von Bildungsreformen und strategischen Maßnahmen der schwedischen Regierung auf den Weg gebracht worden, die auch Bedeutung für den Fremdsprachenunterricht im Fach *Moderna språk* gehabt haben. Anfang der 90er Jahre, als Vorbereitung für den Eintritt in der Europäischen Union, wurde die Notwendigkeit von weiteren Sprachkenntnissen außer Englisch hervorgehoben. Die Bildungsreform 1994 sollte den Fremdsprachenunterricht verstärken. Hierbei wurde auch Spanisch als die dritte optionale Schulsprache eingeführt und den Fremdsprachen Deutsch und Französisch gleichgestellt. Der Ansatz, dass ein höherer Anteil von Schülerinnen und Schülern mit der Reform eine zweite Fremdsprache belegen sollte, hat jedoch anfangs wenig Effekt gehabt (vgl. Tholin 2017). Dahingegen hat dies aber zu einer anderen Veränderung geführt nämlich *welche* Sprachen die Schülerinnen und Schüler in der schwedischen Schule wählen. Noch im Jahr 1996 war Deutsch die beliebteste zweite Fremdsprache in Schweden mit etwa 50 % der schwedischen Schülerinnen und Schülern, die eine zweite Fremdsprache belegten. Zehn Jahre später war aber bereits Spanisch die meistgewählte zweite Fremdsprache in der Grundschule und am Gymnasium und dies ist heute immer noch der Fall (vgl. Kap. 2.2.2).

Die Voraussetzungen für das Erlernen einer Sprache sehen heute für die jeweiligen Fremdsprachen im schwedischen Schulkontext unterschiedlich aus. Englisch hat in der Gesellschaft einen hohen Status (vgl. European Commission 2012a) und wird in vielen Bereichen des Arbeitslebens als notwendig angesehen. Darüber hinaus hat Englisch in den letzten Jahren eine Sonderrolle in Schweden erhalten, da viele Lernende der Sprache täglich auch außerhalb des Unterrichts begegnen. Dies hat dazu geführt, dass Englisch in vielerlei Hinsicht eher als Zweitsprache denn als Fremdsprache betrachtet wird (z. B. Sundquist & Sylvén 2014). Daher liegt der Unterrichtsfokus im Fach traditionell weniger auf den deklarativen Kompetenzen und mehr auf den prozeduralen Kompetenzen

18 Ein Überblick darüber, wie die kommunikative Kompetenz in schwedischen Lehrplänen zwischen den Jahren 1962 und 2000 konzipiert ist und in verschiedenen Lehrbüchern des Deutschen zum Ausdruck kommt, ist in Tornberg (2000) zu finden.

und es ist zu vermuten, dass man beim Erlernen einer weiteren Fremdsprache außer Englisch mehr Zeit investieren muss. Aus diesem Grund sind Maßnahmen, die darauf abzielen, die zweite Fremdsprache in der schwedischen Schule zu stärken, eingeführt und diskutiert worden (vgl. hierzu 2.2.5). Auch wenn manchmal von einer Krise für die zweite Fremdsprache in Schweden gesprochen wird, erscheint es angezeigt zu bemerken, dass insgesamt ein zunehmender Anteil von Schülerinnen und Schülern in der schwedischen Schule eine zweite Fremdsprache belegt (vgl. Krigh 2019; Granfeldt et al. 2021). Das Schulfach *Tyska* hat im gegenwärtigen schwedischen Schulsystem in den letzten Jahren jedoch eine stabile Position als die zweitbeliebteste zweite Fremdsprache mit etwa einem Viertel aller Lernenden, die nach Englisch eine weitere Fremdsprache wählen, gehabt.

2.2 Fremdsprachenlernen in der schwedischen Schule

Das Bildungssystem in Schweden umfasst vier Teile: die Vorschule, Schule, Universitäten und Hochschulen sowie die Erwachsenenbildung. Die einheitliche zehnjährige Grundschule besteht aus einem Vorschuljahr als Vorbereitung und danach insgesamt noch mindestens neun Schuljahren. Noten nach einer sechsgradigen Skala werden ab der 6. Klasse vergeben und mit den Abschlussnoten der 9. Klasse können die Schülerinnen und Schüler sich für ein dreijähriges Gymnasium (entspricht etwa der deutschen Oberstufe) anmelden. Mit der Grundschule endet in Schweden die Schulpflicht und auch wenn die große Mehrheit der schwedischen Jugendlichen ein Gymnasium besucht, ist das Besuchen eines Gymnasiums nicht verpflichtend. Das schwedische Gymnasium besteht aus sowohl theoretischen (studienvorbereitenden) als auch praktischen (berufsvorbereitenden) Ausbildungsprogrammen. Die dreijährige Gymnasialausbildung besteht aus einem System mit unterschiedlichen Kursen. Sowohl die theoretischen als auch die praktischen Gymnasialausrichtungen verlangen eine bestimmte Anzahl an Kursen und die Wahl dieser Kurse ist hauptsächlich durch die Ausrichtung der Ausbildung festgelegt. Mit einer abgeschlossenen Gymnasialausbildung und wenn die für das jeweilige Studium nachgefragten Kurse belegt wurden, können die Lernenden ein Hochschul- oder Universitätsstudium beginnen.

In Schweden sind Schülerinnen und Schüler, wie in vielen anderen europäischen Ländern, verpflichtet, Englisch als erste Fremdsprache zu erlernen (vgl. Broek & van den Ende 2013). Das Schulfach Englisch soll in der Grundschule spätestens ab Klasse 3 unterrichtet werden. Schwedische Grundschulen müssen ihren Schülerinnen und Schülern gemäß schwedischem Bildungsgesetz

zusätzlich auch eine zweite Fremdsprache anbieten (vgl. Utbildningsdepartementet 2010b). Die zweite Fremdsprache wird in der Regel ab Klasse 6 angeboten,[19] ist jedoch zum Teil fakultativ.[20] Dies bedeutet, dass die Schülerinnen und Schüler eine zweite Fremdsprache aus den Sprachen, die von ihrer Grundschule angeboten werden, wählen können. Die Schulen sollen gemäß schwedischem Bildungsgesetz mindestens zwei der modernen Fremdsprachen Deutsch, Französisch und Spanisch anbieten, aber auch andere Fremdsprachen können in Frage kommen.[21] Schwedische Schülerinnen und Schüler haben aber auch andere Optionen: sie können statt einer zweiten Fremdsprache zusätzlichen Unterricht in Schwedisch, Englisch, Schwedisch als Zweitsprache oder Zeichensprache erhalten. Viele Grundschulen bieten daher als Alternative die Fächer Schwedisch/Englisch kombiniert an. Schülerinnen und Schüler mit Migrationshintergrund haben zudem die Möglichkeit, ihre Muttersprache statt einer zweiten Fremdsprache zu wählen (vgl. Utbildningsdepartementet 2011). Eine weitere Fremdsprache kann in der schwedischen Grundschule als Wahlfach belegt werden. Diese Möglichkeit wird an manchen Schulen in der 8. Klasse angeboten.

In der schwedischen Grundschule ist in den letzten Jahren der Anteil der Schülerinnen und Schüler, die eine zweite Fremdsprache lernen, gestiegen. Landesweit beginnen etwa 90 % aller Schülerinnen und Schüler mit einer zweiten Fremdsprache. In der 9. Klasse sind etwa 77 % der Schülerinnen und etwa

---

19 Früher haben Schülerinnen und Schüler der schwedischen Grundschule entweder ab der 6. oder 7. Klasse mit ihrer zweiten Fremdsprache begonnen. Seit dem Schuljahr 2018/2019 sollen schwedische Schülerinnen und Schüler mit der zweiten Fremdsprache (meistens Deutsch, Französisch oder Spanisch) in Klasse 4–6 beginnen, dementsprechend spätestens ab der 6. Klasse der Grundschule. Dies hat aber dazu geführt, dass einige Schulen, die vorher drei moderne Sprachen angeboten haben, nur noch zwei moderne Sprachen zur Wahl stellen (vgl. Bardel et al. 2019).

20 Die Wahl einer zweiten Fremdsprache in Schweden ist nicht obligatorisch, wird aber von der schwedischen Schulbehörde stark gefördert (Skolverket 2000). Obligatorisch ist hingegen das Schulfach „Sprachwahl" (*språkval* – 320 Stunden im schwedischen Lehrplan), worin schwedische Lernende zurzeit neben einer zweiten Fremdsprache auch andere Optionen haben. Dies wird jedoch voraussichtlich geändert, um u. a. mehr Fokus auf die zweite Fremdsprache zu richten.

21 Die große Mehrheit der schwedischen Grundschulen bietet alle drei Schulsprachen Deutsch, Französisch und Spanisch an (Granfeldt et al. 2019a). Außer diesen am häufigsten angebotenen modernen Sprachen können an gewissen Schulen auch weitere Fremdsprachen wie Dänisch, Chinesisch, Italienisch, Japanisch und Russisch gewählt werden. Für Chinesisch existieren separate Bildungsstandards.

70 % der Schüler bei ihrer Sprachwahl geblieben (vgl. Skolverket 2021c). Im Vergleich mit entsprechenden Zahlen aus den Jahren 1997–2010 (vgl. Tholin 2017) zeigt dies eine leichte Erhöhung. Im Einklang damit hat auch der Anteil der Lernenden, der am Gymnasium das Fach *Moderna språk* belegt, zugenommen. Diese Erhöhung ist eventuell auf nationale Maßnahmen im Jahr 2007 zurückzuführen, die das Interesse für Fremdsprachen erhöhen sollten, z. B. das Einführen von Leistungspunkten, sog. Meritpunkten (*meritpoäng*), (vgl. Granfeldt et al. 2021, siehe auch Kap. 2.1.4).[22] Jedoch ist die immer noch relativ hohe Abwahlquote der Sprachlernenden bis zur 9. Klasse, vor allem unter Jungen (vgl. Cardelús 2015: 163), zu bedenken.[23]

Am weiterführenden Gymnasium können die Schülerinnen und Schüler landesweit zwischen 18 Studienprogrammen, sechs studienvorbereitenden[24] und 12 berufsvorbereitenden[25], wählen.[26] In der Gymnasialschule kann die in der Grundschule gewählte zweite Fremdsprache fortgesetzt werden oder mit einer neuen Fremdsprache begonnen werden. Bei der großen Mehrheit der achtzehn verschiedenen Studienprogramme in Schweden ist eine Fremdsprache außer Englisch jedoch kein Pflichtfach. Eine zweite Fremdsprache ist lediglich in vier der sechs studienvorbereitenden Studiengängen obligatorisch. Es handelt sich dabei um theoretische Studienprogramme, die den Zugang

22 Diese Zunahme hat aber vor allem in den städtischen Regionen stattgefunden. Auch wenn das Fach Deutsch aber in etwa höherem Ausmaß in ländlichen Regionen und mittelgroßen Städten gewählt wird, kann auch für das Fach Deutsch eine leichte Erhöhung wahrgenommen werden (vgl. Granfeldt et al. 2021).

23 Das Wählen oder Abwählen einer zweiten Fremdsprache in Schweden scheint aber stark mit familiärem Hintergrund und Geschlechtsunterschieden zusammenzuhängen: Mädchen lernen in höherem Ausmaß als Jungen eine zweite Fremdsprache und Kinder in sozioökonomisch schwächeren Gruppen neigen eher dazu, eine zweite Fremdsprache abzuwählen (vgl. Krigh 2019).

24 Die studienvorbereitenden Programme sind folgende: Ästhetisches Programm, Geisteswissenschaftliches Programm, Gesellschaftswissenschaftliches Programm, Naturwissenschaftliches Programm, Technisches Programm und Wirtschaftliches Programm.

25 Als berufsvorbereitende Programme zählen z. B. Bau- und Anlagenprogramm, Gesundheitsfürsorge- und Pflegeprogramm, Handels- und Verwaltungsprogramm, Handwerksprogramm, Hotel- und Tourismusprogramm, Industrietechnisches Programm, Kinder- und Freizeitbetreuungsprogramm, sowie Restaurant- und Lebensmittelprogramm.

26 Darüber hinaus gibt es neben den regulären Studienprogrammen zusätzlich auch lokale Spezialprogramme, die landesweit gewählt werden können.

zur Hochschulausbildung und zum Universitätsstudium gestatten. Das Belegen einer zweiten Fremdsprache umfasst beispielsweise bei einem Programm naturwissenschaftlicher Ausrichtung einen Kurs mit 100 Punkten, d. h. mindestens eine Sprachstufe (etwa ein Schuljahr), während die geistes- und sozialwissenschaftlichen Ausrichtungen mindestens zwei Kurse mit 200 Punkten, d. h. zwei Sprachstufen (etwa zwei Schuljahre), verlangen. Weitere Sprachstufen können als Wahlfach belegt werden. In Schweden gibt es für einige Gymnasialprogramme die Möglichkeit, zusätzlich eine dritte und vierte Fremdsprache aus dem Sprachangebot der jeweiligen Schule zu wählen.[27] Bei den berufsvorbereitenden Gymnasialprogrammen ist eine zweite Fremdsprache lediglich fakultativ.

Schwedische Schülerinnen und Schüler am Gymnasium können demzufolge, abhängig vom Sprachangebot ihrer Schule, auf ihre Sprachkenntnisse von der Grundschule aufbauen, aber haben auch die Möglichkeit, eine neue Sprache zu erlernen. Daher sollen gemäß dem Bildungsgesetz für die Gymnasialschule die modernen Sprachen Deutsch, Französisch und Spanisch immer angeboten werden, sowohl auf einem Anfängerniveau als auch auf einem fortgeschrittenen Niveau, das auf die Sprachkenntnisse aus der Grundschule aufbaut.[28] Die Schulen können auch weitere Fremdsprachen oder Zeichensprache anbieten. Wie in der Grundschule können Schülerinnen und Schüler mit Migrationshintergrund Unterricht in der Muttersprache anstatt in einer zweiten Fremdsprache wählen (vgl. Utbildningsdepartementet 2010a).

2.2.1 Einheitliches System für den Fremdsprachenunterricht

Im Jahr 2000 wurde in Schweden ein einheitliches System im Hinblick auf Englisch und die modernen Sprachen mit einer für die Grund- und Gymnasialschule gemeinsamen Progression in sieben Niveaustufen eingeführt. Diese

---

27 Hier handelt es sich vor allem um Studienprogramme, die eine sprachliche Spezialisierung haben, wie das Geisteswissenschaftliche Programm.

28 Diese Formulierung scheint jedoch von den Schulen unterschiedlich interpretiert zu werden (vgl. Skolverket 2018a). Manche Gymnasialschulen interpretieren die Formulierung als einen Hinweis darauf, dass lediglich Unterricht der Fremdsprachenniveaus 1 und 3 verpflichtend ist. Dies bedeutet u. a., dass Schülerinnen und Schüler, die eine Fremdsprache in der Grundschule als Wahlfach belegt haben, d. h. Stufe 1 belegt haben, oder am Gymnasium eine neue Sprache gewählt haben, ebenfalls Stufe 1, mit ihrer Fremdsprache auf Stufe 2 nicht immer fortfahren können. Andere aber interpretieren die Formulierung als einen Hinweis darauf, dass Sprachunterricht in Deutsch, Französisch und Spanisch lediglich für theoretische Gymnasialprogramme verpflichtend ist.

Niveaustufen wurden im schwedischen System als *steg*[29] bezeichnet. Wenn eine zweite Fremdsprache von der 6. Klasse bis zur 9. Klasse der Grundschule belegt wird, entspricht dies Stufe 1 und Stufe 2 im schwedischen System für Fremdsprachen. Diese Schülerinnen und Schüler können am Gymnasium auf Stufe 3 ihre in der Grundschule gewählte Sprache weiterlernen, vgl. die erste Alternative in Abb. 1:

Stufe 1 Sprachwahl Gr 6–7	Stufe 2 Sprachwahl Gr 8–9	Stufe 3 Gy	Stufe 4 Gy	Stufe 5 Gy	Stufe 6 Gy	Stufe 7 Gy
Stufe 1 Wahlfach Gr 8–9	Stufe 2 Gy	Stufe 3 Gy	Stufe 4 Gy	Stufe 5 Gy	Stufe 6 Gy	Stufe 7 Gy
Stufe 1 Gy	Stufe 2 Gy	Stufe 3 Gy	Stufe 4 Gy	Stufe 5 Gy	Stufe 6 Gy	Stufe 7 Gy

Abb. 1: Überblick über die sieben Sprachenniveaus für Moderna språk im schwedischen Bildungssystem für die Grund- (Gr) und Gymnasialschule (Gy) (nach Skolverket 2000)

Schülerinnen und Schüler, die in der 8. Klasse zusätzlich eine neue Sprache erlernen, erreichen am Ende der Grundschule in der 9. Klasse die erste Niveaustufe. In der Gymnasialschule kann es die Möglichkeit geben, diese in der Grundschule gewählte zusätzliche Fremdsprache auf Stufe 2 weiter zu belegen, vgl. die zweite Alternative in Abb. 1 oben. Am Gymnasium kann aber auch mit einer neuen Fremdsprache auf Stufe 1 begonnen werden, vgl. die dritte Alternative in Abb. 1. In der Grundschule wird demzufolge eine Stufe nach zwei Jahren erreicht, während am Gymnasium der Verlauf schneller ist und eine Stufe daher nach einem Jahr abgeschlossen wird.

Eine Mehrheit der Lernenden in studienvorbereitenden Gymnasialausrichtungen führen die bei der Sprachwahl in der Grundschule gewählte Sprache fort und belegen die dritte und vierte Stufe in ihrer Sprache. Maximal können schwedische Schülerinnen und Schüler die Niveaustufe 7 erreichen. Die Niveaustufe 5 wird allerdings von einer geringeren Anzahl von Lernenden belegt und die beiden Niveaustufen 6–7 kommen im schwedischen Bildungssystem selten vor. Wenn Schülerinnen und Schüler am Gymnasium eine weitere Fremdsprache wählen, beginnen sie mit der ersten Niveaustufe.

---

29 Die Bezeichnungen *steg 1*, *steg 2* („Stufe 1“ bzw. „Stufe 2“), usw. wurden jedoch mittlerweile von Skolverket durch *Moderna språk 1*, *Moderna språk 2* („Moderne Sprache 1“ bzw. „Moderne Sprache 2“), usw. ersetzt. Um Missverständnisse zu vermeiden werden in der vorliegenden Arbeit weiterhin die Bezeichnungen „Stufe“, „Sprachstufe“ oder „Fremdsprachenstufe“ verwendet.

2.2.2 Die aktuelle Stellung des Faches Deutsch in der schwedischen Schule

Traditionell werden, wie bereits erwähnt, in Schweden Deutsch und Französisch als zweite Fremdsprache angeboten, aber seit der Schulreform 1994 zählt nun auch Spanisch zu den zentralen Sprachen. Da die Grundschulen verpflichtet sind, mindestens zwei der Sprachen Deutsch, Französisch und Spanisch anzubieten, werden andere Sprachen selten zur Wahl gestellt. Spanisch ist zurzeit in der schwedischen Grundschule die meistgewählte Sprache, gefolgt von Deutsch und danach Französisch. Innerhalb der Sprachwahl *Moderna språk* für das Schuljahr 2019/2020 lernten in der 9. Klasse der Grundschule etwa 57 % der Schülerinnen und Schüler Spanisch, etwa 24 % Deutsch und etwa 18 % Französisch. Lediglich 0,4 % hatten andere Sprachen belegt, vorwiegend Chinesisch, Finnisch, Samisch oder Arabisch.[30] In der Gymnasialschule können, je nach Angebot der Schule, weitere Fremdsprachen gelernt werden. Die etablierten Schulsprachen Deutsch, Französisch und Spanisch sind auch am Gymnasium die meistgewählten Fremdsprachen, wie folgende Statistik der schwedischen Schulbehörde aus dem Schuljahr 2019/2020 zeigt, siehe Tab. 1:

Tab. 1: *Verteilung schwedischer Lernender am Gymnasium mit einer Abschlussnote im Fach* Moderna språk *nach gewählten Sprachen, Schuljahr 2019/2020*

Sprache	*Deutsch*	*Französisch*	*Spanisch*	*Italienisch*	*Sonstige Sprachen*
Anzahl	11 758	9 401	24 897	2 292	2 803
Prozent	23 %	18 %	49 %	5 %	5 %

Die große Mehrheit der schwedischen Schülerinnen und Schüler, die am Gymnasium eine zweite Fremdsprache belegen, hat dementsprechend die Sprachen Spanisch, Deutsch und Französisch gewählt, was natürlich auch vom Angebot der Gymnasialschulen abhängig ist (vgl. Granfeldt et al. 2019a). Gemäß Statistik für das Schuljahr 2019/2020 erhielten über 11 700 Lernende am Gymnasium eine Abschlussnote im Fach Deutsch. Im Fach Spanisch sind es etwas mehr als doppelt so viele, 2020 erhielten etwa 24 900 Schülerinnen und

30 Eigene Bearbeitung personenbezogener Statistiken gesammelt von *Swedish Statistics* (SCB) und durch *Skolverket* zur Verfügung gestellt. Diese Datenbank, die u. a. Informationen über die Anzahl der Schülerinnen und Schüler im Fach *Moderna språk* nach gewählten Sprachen in Schweden enthält, wird auch im Folgenden verwendet.

Schüler eine Abschlussnote in Spanisch. Bei den weiteren Fremdsprachen liegt Französisch mit 9 400 Lernenden vor den etwa 2 300 Lernenden in Italienisch. Eine geringere Anzahl von Schülerinnen und Schülern erhielt eine Abschlussnote in sonstigen Fremdsprachen wie Arabisch, Dänisch, Chinesisch, Japanisch und Russisch. Diese Verteilung zwischen den jeweiligen Fremdsprachen ist in den letzten Jahren am schwedischen Gymnasium relativ unverändert geblieben.

In der Gymnasialschule setzt die Mehrheit der Schülerinnen und Schüler das Studium ihrer in der Grundschule erlernten Sprache fort. Dies bedeutet meistens, dass sie im ersten Jahr am Gymnasium mit dem dritten Fremdsprachenniveau (z. B. *Tyska 3*) beginnen. Mehr als die Hälfte belegen im Folgejahr auch die vierte Fremdsprachenstufe (z. B. *Tyska 4*), während deutlich weniger Lernende mit der fünften Fremdsprachenstufe (z. B. *Tyska 5*) weitermachen. Dies wird auch in der Verteilung auf die Fremdsprachenstufen für das Fach Deutsch ersichtlich. Tab. 2 gibt einen aktuellen Überblick über die Verteilung der Deutschlernenden am schwedischen Gymnasium auf die jeweiligen Fremdsprachenstufen in den vergangenen Jahren:

Tab. 2: *Anzahl schwedischer Schülerinnen und Schüler mit einer Note im Fach Deutsch am Gymnasium, pro Kurs (1–7) und Schuljahr*

Schuljahr	*2015/16*	*2016/17*	*2017/18*	*2018/19*	*2019/20*
Tyska 1	2 983	3 190	3 049	2 740	2 891
Tyska 2	1 426	1 429	1 371	1 216	1 200
Tyska 3	7 704	7 876	7 982	8 733	8 937
Tyska 4	5 037	5 097	5 088	5 457	5 511
Tyska 5	540	493	459	451	348
Tyska 6	84	77	62	76	58
Tyska 7	50	45	39	46	40

Zu erkennen ist, dass die Verteilung der Deutschlernenden pro Sprachniveau über die letzten fünf Jahre hinweg relativ stabil ist, auch wenn die Zahlen von Jahr zu Jahr schwanken. In den letzten Jahren haben pro Jahr etwa 3 000 schwedische Jugendliche am Gymnasium Deutsch als Anfängersprache, *Tyska 1*, belegt. Es handelt sich dabei um Lernende ohne Vorkenntnisse, die von der Grundschule keine Abschlussnote in Deutsch haben und entweder eine andere Fremdsprache belegten oder einen verstärkten Schwedisch/Englischunterricht erhielten. Ungefähr die Hälfte dieser Anzahl, etwa 1 200 bis 1 400 Schülerinnen und Schüler, wählen pro Jahr den auf *Tyska 1* aufbauenden Kurs *Tyska 2*. Diese Deutschlernenden haben entweder in der 8. Klasse der Grundschule Deutsch als Sprachwahl gehabt oder *Tyska 1* am Gymnasium belegt. *Tyska 2* wird von

manchen Schulen aufgrund von Sparbeschlüssen erst dann angeboten, wenn die Schülergruppen genügend Teilnehmer haben. Der etwas negative Trend, die Kurse *Tyska 1* bzw. *Tyska 2* zu absolvieren, könnte zudem daran liegen, dass immer mehr Lernende, eventuell aufgrund des Systems mit Meritpunkten, ihre in der Grundschule gewählte Fremdsprache weiterlernen.

Wahrscheinlich auf Grund der Meritpunkte kann eine leichte Erhöhung in den Fremdsprachenstufen *Tyska 3* und *Tyska 4* wahrgenommen werden. In der Regel wählt die Mehrheit der Lernenden am Gymnasium erneut ihre in der Grundschule gewählte Sprache und fängt mit der dritten Fremdsprachenstufe an. Die Kurse, in denen die zweite Fremdsprache aus der Grundschule fortgeführt wird, sind auch für das Fach Deutsch die meistgewählten Sprachstufen: eine steigende Anzahl, im Jahr 2020 fast 9 000 Schülerinnen und Schüler auf *Tyska 3* und etwa 5 500 auf *Tyska 4*, belegen diese Stufen. Die geringere Anzahl von Deutschlernenden auf *Tyska 4* im Vergleich zu *Tyska 3* hängt wahrscheinlich damit zusammen, dass die naturwissenschaftliche Gymnasialausrichtung das Belegen einer zweiten Fremdsprache nur im Umfang eines Kurses verlangt. Der Kurs *Tyska 4* kann aber von interessierten Schülerinnen und Schülern innerhalb dieser Ausrichtung als Wahlfach belegt werden. Da die fünfte Stufe, *Tyska 5*, nicht an allen Gymnasialschulen angeboten wird oder jedes Jahr zustande kommt, wird diese Stufe von einer deutlich geringeren Anzahl von Lernenden belegt. Der Trend zeigt, dass immer weniger Schülerinnen und Schüler die höheren Stufen in Deutsch belegen, im Schuljahr 2019/2020 haben nur etwa 350 Lernenden eine Abschlussnote im Kurs *Tyska 5* erhalten.[31] Um genügend Schülerinnen und Schüler pro Fremdsprache und Stufe zu erhalten, kooperieren manche Schulen und organisieren für eine oder mehrere Sprachstufen gemeinsamen Unterricht. Einige Schulen bieten auch Fremdsprachenunterricht in gemischten Gruppen, d. h. mit zwei oder sogar mehreren Sprachstufen, an.

2.2.3 Schriftliche Kompetenz in schwedischen Lehrplänen

Die schwedischen Lehrpläne für die Fremdsprachen enthalten Beschreibungen bestimmter Inhaltsbereiche (*content standards*), die im Unterricht behandelt werden sollen, sowie Anforderungen (*performance standards*), die verschiedene Leistungsniveaus definieren. Parallel mit einem verstärkten Fokus auf

31 Die Anzahl der Lernenden in den höheren Kursen *Tyska 6* und *Tyska 7*, die ebenfalls wie *Tyska 5* nur von wenigen Schulen angeboten werden, liegen jeweils unter 100 Lernenden (in etwa 60 bzw. 40 Lernende im Jahr 2020) pro Schuljahr.

Lernergebnisse und mit dem Einführen von international anerkannten Leistungsmessungsstudien haben die Lehrpläne in gegenwärtigen Reformentwicklungen der schwedischen Schule eine tragende Bedeutung erhalten. Diese Lernergebnisorientierung, häufig als sog. *outcome-based-education (OBE)* bezeichnet, kann in den heutigen schwedischen Lehrplänen für *Moderna språk* aus dem Jahr 2011 (vgl. Skolverket 2011a[32]) beobachten werden, auch im Vergleich zu den anderen nordischen Ländern (Wahlström 2016: 90 ff.). Kennzeichnend für diese Lehrpläne sind Erwartungen, in denen deutliche Anforderungen formuliert werden, die Schülerinnen und Schüler am Ende einer Lerneinheit erfüllt haben sollten. Definiert sind hierbei auch Mindestanforderungen zu jeder Lerneinheit, deren Bewältigung von allen Schülerinnen und Schülern dieser Lerneinheit erwartet wird. Des Weiteren werden Tests und Lernergebnisse in resultatorientierten Ansätzen häufig mit Standards international anerkannter Rahmenwerke, wie des Referenzrahmens für Sprachen, in Verbindung gesetzt (vgl. Chapelle 2020).

Die heutigen schwedischen Bildungsstandards für *Moderna språk* folgen einem für Grundschule und Gymnasium gemeinsamen System mit sieben Kompetenzstufen, die aufeinander aufbauen und die sich an dem GER orientieren (vgl. Kap. 2.3.2). Die Lehrpläne jener Kompetenzstufen stellen Mindestanforderungen, die Schülerinnen und Schüler nach einer Lerneinheit erfüllt haben sollten. Für jede Stufe sind in den heutigen Bildungsstandards für das Erlernen der zweiten Fremdsprache bestimmte Mindestanforderungen formuliert. Dies schließt allerdings nicht aus, dass Schülerinnen und Schüler innerhalb der jeweiligen Fremdsprachenstufen auch höhere Kompetenzen zeigen können und auch Beschreibungen höherer Anforderungen sind dementsprechend in den Lehrplänen der jeweiligen Kompetenzstufen zu finden. Die Bildungsstandards für *Moderna språk* sind in drei Teile unterteilt:

1. Sinn und Zweck des Faches
2. zentraler Inhalt
3. Wissensanforderungen.

Anfangs, im *Sinn und Zweck des Faches* (*ämnets syfte*), benennen die Bildungsstandards für *Moderna språk* die Bedeutung des Faches für die erhöhten Möglichkeiten jedes einzelnen Individuums zu sozialen und kulturellen Kontakten

32 Aktuell ist eine Überarbeitung der schwedischen Bildungsstandards aus dem Jahr 2011 im Hinblick auf die Fremdsprachen in Schweden durchgeführt worden. Die überarbeitete Version für das Gymnasium gilt seit 1. Juli 2021 (vgl. Skolverket 2021a).

und einem erweiterten Verständnis für das Leben anderer Menschen sowie übergreifende fachspezifische Ziele und Richtlinien für den Unterricht. Der Fokus liegt hierbei auf der kommunikativen Funktion der Sprache: Es wird deutlich angestrebt, dass die Lernenden durch den Sprachgebrauch in funktionalen und sinnvollen Kontexten eine vielseitige Kommunikationsfähigkeit entwickeln. Diese Fähigkeit umfasst für den Kompetenzbereich Schreiben, d. h. die schriftliche Produktion und Interaktion, sich in der Fremdsprache in Schrift ausdrücken und mit anderen interagieren zu können sowie situations- und partneradäquate Texte zu schreiben (vgl. Skolverket 2011a). Die Sprachverwendung steht wie im Referenzrahmen deutlich im Vordergrund. Die Bildungsstandards sind gegen den internationalen Trend nicht deutlich kompetenzorientiert (vgl. Wahlström 2016: 94) und verzichten damit auch auf eine Einteilung der kommunikativen Kompetenz in Teilkompetenzen, wie z. B. linguistische, soziolinguistische oder pragmatische Kompetenzen (vgl. Europarat 2001).

Des Weiteren wird im *zentralen Inhalt* (*centralt innehåll*) das beschrieben, was im Unterricht behandelt werden soll. Diese Beschreibungen sind in den heutigen Lehrplänen detaillierter dargestellt als in den bisherigen Lgr 80 und Lpo 94 / Lpf 94 (Wahlström 2016: 93). Der zentrale Inhalt ist in den einzelnen Kursbeschreibungen jeder Fremdsprachenstufe zu finden und wird zunächst in folgende drei Bereiche gegliedert: *Kommunikationsinhalt*, *Rezeption* sowie *Produktion und Interaktion*. Die im Mittelpunkt der vorliegenden Arbeit stehende Teilkompetenz, die schriftliche Kompetenz, ist im Lehrplan unter *Produktion und Interaktion* zu finden. Die zentralen Inhalte hinsichtlich der Produktion und Interaktion auf den untersuchten Fremdsprachenstufen *Tyska 3*, *Tyska 4* und *Tyska 5* sind in Tab. 3 aufgeführt.[33]

Die Tabelle veranschaulicht Lernstoff und Aktivitäten, die als Ausgangspunkt für den Unterricht in der zweiten Fremdsprache hinsichtlich Interaktion und Produktion dienen sollen. Hierbei geht es z. B. um Strategien, um sprachliche Probleme zu lösen, Bearbeitungen eigener und fremder Textproduktionen vornehmen zu können sowie Texte angegebener Textsorten mit zunehmender sprachlicher Sicherheit zu verfassen. In den Lehrplänen werden explizit konkrete Sprachhandlungen und Kontexte angegeben, die die Lernenden mündlich und schriftlich bewältigen sollen. Es geht u. a. um instruierende, narrative

33 Auszug im Original im Anhang 1 (*eigene Übersetzung, M.H.R.*).

Tab. 3: *Zentrale Inhalt hinsichtlich Produktion und Interaktion in den schwedischen Bildungsstandards für Tyska 3, Tyska 4 und Tyska 5 (Skolverket 2011a)*

Tyska 3	*Tyska 4*	*Tyska 5*
Anleitungen, Erzählungen und Beschreibungen in zusammenhängendem Sprechen und Schreiben. Diskussionen, Gespräche und Schreiben für Kontakt und Kommunikation in verschiedenen Situationen.	Anleitungen, Erzählungen und Beschreibungen in zusammenhängendem Sprechen und Schreiben. Gespräche, Diskussionen, und Argumentation für Kommunikation und Kontakt in verschiedenen Situationen.	Mündliche und schriftliche Produktion und Interaktion verschiedener Art, auch in formelleren Kontexten, wo die SchülerInnen instruieren, erzählen, zusammenfassen, erklären, kommentieren, bewerten, ihre Meinungen begründen, diskutieren und argumentieren.
Strategien, um sprachliche Probleme zu lösen, z. B. mithilfe von Umformulierungen und Erklärungen	Strategien, um sprachliche Probleme zu lösen, z. B. mithilfe von Umformulierungen, Fragen und Erklärungen.	
Strategien, um zu Gesprächen beizutragen und aktiv teilzunehmen, z. B. indem man Initiative zur Interaktion ergreift, aktiv zuhört und höflich endet.	Strategien, um zu Gesprächen beizutragen und aktiv teilzunehmen, z. B. indem man Bestätigung gibt, Rückfragen stellt und Initiative zu neuen Fragestellungen und Themenbereichen ergreift.	Strategien, um zu Diskussionen in Bezug auf Gesellschaft und Arbeitsleben beizutragen und aktiv teilzunehmen.
Sprachliche Sicherheit z. B. in Bezug auf Aussprache, Intonation, idiomatische Ausdrücke und grammatische Strukturen in Richtung Deutlichkeit, Variation und Anpassung an Ziel, Partner und Situation.	Sprachliche Sicherheit z. B. in Bezug auf Aussprache, Intonation, idiomatische Ausdrücke und Satzbau in Richtung Deutlichkeit, Variation und Flüssigkeit.	
	Bearbeitung eigener und fremder mündlicher und schriftlicher Produktionen, um diese zu variieren, zu verdeutlichen, zu spezifizieren und an Ziel, Partner und Situation anzupassen.	Bearbeitungen eigener und fremder mündlicher und schriftlicher Produktionen, um diese zu variieren, zu verdeutlichen, zu spezifizieren sowie Struktur zu verschaffen und an Ziel, Partner und Situation anzupassen. Dies beinhaltet die Verwendung von Wörtern und Phrasen, die Kausalzusammenhänge und Zeitaspekte verdeutlichen.

und beschreibende Sprachhandlungen, die mit zunehmender Fremdsprachenstufe einen höheren Komplexitätsgrad erhalten.

Für die höhere Stufe *Tyska 4* werden die Bildungsstandards beispielsweise auch mit argumentierender Kommunikation erweitert und auf *Tyska 5* sollen die Lernenden zudem u. a. Kommunikation in formellen Situationen durchführen und ihre eigene Meinung begründen können. Die Schülerinnen und Schüler sollen mit zunehmender Komplexität außerdem Strategien entwickelt haben, um sprachliche Probleme zu lösen und um aktiv zu Diskussionen beizutragen. Am Ende der Fremdsprachenstufen *Tyska 3* und *Tyska 4* sollen sie eine sprachliche Sicherheit hinsichtlich Aussprache, Intonation, idiomatischen Ausdrücken und grammatischen Strukturen entwickelt haben. Während die sprachliche Sicherheit auf *Tyska 3* für die Variation und die Deutlichkeit eine Bedeutung hat, ist sie auf *Tyska 4* auch im Hinblick auf die Flüssigkeit relevant. Dazu sollen die Lernenden am Ende der Fremdsprachenstufen *Tyska 4* und *Tyska 5* fähig sein, Bearbeitungen eigener und anderer Leistungen durchzuführen, um diese in vielerlei Hinsicht zu verbessern, u. a. im Hinblick darauf, die eigenen Leistungen ziel-, situations- und partneradäquat anzupassen (Skolverket 2011a).[34]

Darüber hinaus umfassen die Lehrpläne der jeweiligen Fremdsprachenstufen auch *Wissensanforderungen* (*kunskapskrav*) für die Noten E, C und A. Die Schülerinnen und Schüler müssen danach bestimmte Fertigkeiten hinsichtlich Rezeption, Produktion und Interaktion zeigen können. Die Wissensanforderungen bestimmen nicht nur die Mindestforderungen dafür, wann ein Lernender den Kurs bestanden hat, sondern auch wann ein Lernender den Kurs gut oder sehr gut bestanden hat, wobei der Komplexitätsgrad sich mit der Note erhöht. Bestimmte Kriterien sind folglich für das Mindestniveau, Note E, und für die höheren Notenstufen C und A formuliert worden. Die Vergabe der Zwischennoten D und B wird dann aktuell, wenn die Wissensanforderungen für

34 In der überarbeiteten Version aus dem Jahr 2021 (vgl. Skolverket 2021a) werden im zentralen Inhalt die Strategien im Hinblick auf die *Produktion und Interaktion* zusätzlich spezifiziert: es handelt sich z. B. um Strategien, um zu Gesprächen und schriftlicher Interaktion (auch digital) beizutragen und sie erleichtern zu können. Zu den Veränderungen gehören zudem dahingehende Änderungen, dass die Anforderungen hinsichtlich der mündlichen und der schriftlichen Sprachfertigkeit getrennt stehen. Zu erwähnen ist aber auch, dass die Formulierungen bezüglich Bearbeitungen und Verbesserungen der eigenen Produktion von den Wissensanforderungen zum zentralen Inhalt gezogen wurden und bereits bei *Tyska 3* aufgeführt werden (ibid.). Da zurzeit der Datenerhebung die Bildungsstandards für die zweite Fremdsprache aus dem Jahr 2011 aktuell waren, werden diese hier und im Folgenden verwendet.

die niedrige Stufe (d. h. die Noten E bzw. C) erfüllt sind und der/die Lernende gleichzeitig mehr als die Hälfte der Wissensanforderungen für die höhere Stufe (d. h. die Noten C bzw. A) erreicht hat. Wenn Lernende am Ende des Kurses die Wissensanforderungen für das unterste Niveau (Note E) nicht erfüllen, erhalten sie eine nicht ausreichende Note (Note F).

Um einen Überblick über die Mindestanforderungen der jeweiligen Fremdsprachenstufen zu verschaffen, werden die Wissensanforderungen für die Note E im Hinblick auf die Produktion und die Interaktion auf den in der

Tab. 4: *Mindestkriterien hinsichtlich Produktion und Interaktion in den schwedischen Bildungsstandards für Tyska 3, Tyska 4 und Tyska 5 (Skolverket 2011a)*

Tyska 3	*Tyska 4*	*Tyska 5*
In mündlichen und schriftlichen Produktionen verschiedener Art formuliert der/die SchülerIn **einfach, verständlich und teilweise** zusammenhängend. Um die eigene Kommunikation zu verdeutlichen und zu variieren, bearbeitet der/die SchülerIn seine/ihre eigenen Produktionen und macht **einfache** Verbesserungen.	In mündlichen und schriftlichen Produktionen verschiedener Genres formuliert der/die SchülerIn **einfach, verständlich und relativ** zusammenhängend. Um die eigene Kommunikation zu verdeutlichen und zu variieren, bearbeitet der/die SchülerIn seine/ihre eigenen Produktionen und macht **einfache** Verbesserungen.	In mündlichen und schriftlichen Produktionen verschiedener Genres formuliert der/die SchülerIn **relativ** variiert, **relativ** deutlich und **relativ** zusammenhängend. Der/die SchülerIn formuliert auch mit **gewisser** Flüssigkeit und **zum Teil** ziel-, partner- und situationsadäquat. Der/die SchülerIn bearbeitet seine/ihre eigenen Produktionen und macht **einfache** Verbesserungen.
In mündlicher und schriftlicher Interaktion formuliert der/die SchülerIn **verständlich und einfach**. Darüber hinaus wählt und verwendet der/die SchülerIn **hauptsächlich** funktionierende Strategien, die **zum Teil** Probleme lösen und die Interaktion verbessern.	In mündlicher und schriftlicher Interaktion verschiedener Art formuliert der/die SchülerIn **verständlich und einfach sowie zum Teil** ziel-, partner- und situationsadäquat. Darüber hinaus wählt und verwendet der/die SchülerIn **hauptsächlich** funktionierende Strategien, die **zum Teil** Probleme lösen und die Interaktion verbessern.	In mündlicher und schriftlicher Interaktion verschiedener Art, auch in formelleren Kontexten, formuliert der/die SchülerIn deutlich und mit **gewisser** Flüssigkeit sowie **zu gewissem Grad** ziel-, partner- und situationsadäquat. Darüber hinaus wählt und verwendet der/die SchülerIn **hauptsächlich** funktionierende Strategien, die **zum Teil** Probleme lösen und die Interaktion verbessern.

vorliegenden Arbeit untersuchten Fremdsprachenstufen *Tyska 3*, *Tyska 4* und *Tyska 5* in Tab. 4 zusammengefasst.[35]

Die Wissensanforderungen hinsichtlich schriftlicher (und mündlicher) Produktion und Interaktion auf den jeweiligen Stufen zeigen häufig eine Progression zwischen den Fremdsprachenstufen. Es geht hierbei um das Verfassen von Texten verschiedener Art, was auf den höheren Stufen auch bedeutet, dass die Schülerinnen und Schüler Texte verschiedener Genres verfassen können. Des Weiteren sollen die Lernenden einfach und verständlich formulieren können, auf höheren Stufen zudem mit einem höheren Komplexitätsgrad hinsichtlich Kohäsion, Variation und Deutlichkeit. Auch Anforderungen hinsichtlich soziolinguistischer Kompetenz, z. B. Texte ziel-, partner- und situationsbezogen zu gestalten, sollen auf den höheren Stufen erfüllt werden. Die Kompetenz des schriftlichen Ausdrucks wird im Lehrplan als Prozess betrachtet, indem explizit angestrebt wird, dass die Lernenden ihre eigenen Textproduktionen bearbeiten sollen und dabei einfache Verbesserungen leisten können. Darüber hinaus sollen die Schülerinnen und Schüler auf diesen Stufen funktionierende Strategien verwenden können, um sprachliche Probleme zu lösen und die Interaktion zu verbessern (Skolverket 2011a).

Insgesamt zeigen die Wissensanforderungen für die zweite Fremdsprache in Schweden, wie der europäische Referenzrahmen, auf einen deutlich handlungs- und kompetenzorientierten Ansatz. Die Bewertungskriterien hängen somit auch mit der Zielsetzung des Faches und dem zentralen Inhalt eng zusammen. Auch wenn die schriftliche Interaktion und Produktion einen zunehmenden Grad an sprachlicher Sicherheit enthalten soll, steht die sprachliche Form an sich nicht im Mittelpunkt. Betont wird eher das, wozu die Lernenden ihre Sprache verwenden können, wie z. B. Fragen stellen, über etwas erzählen können und ihre Meinungen ausdrücken. Obwohl die kommunikative Funktion der Sprache, wie im GER, im Vordergrund steht, sind die schwedischen Standards aber allgemeiner als der GER formuliert. Aus diesem Grund könnte ein textueller Vergleich mit den eher detaillierten Deskriptoren und Skalen des GER schwerer fallen (vgl. hierzu Oscarson 2015).

2.2.4 Bewertung und fakultative Tests der zweiten Fremdsprache

In Schweden gibt es am Ende der Grund- oder Gymnasialschule keine besonderen Abschlussprüfungen. Schwedische Schülerinnen und Schüler erhalten

---

35 Auszug im Original im Anhang 2 (*eigene Übersetzung, M.H.R., Hervorheb. im Original*).

aber am Ende der 9. bzw. 12. Jahrgangsstufe ein Abschlusszeugnis, das für den Zugang zum Gymnasium bzw. zu höheren Studien nötig ist. Die Leistungsbeurteilung geschieht generell durch die praktizierenden Lehrkräfte, basierend auf Dokumentationen aus dem Unterrichtsalltag wie Klausuren, Aufgaben oder anderen Aktivitäten. Dazu sind in einigen Schulfächern landesweite Leistungstests (*nationella prov*) vorhanden (z. B. in Englisch, Mathematik und Schwedisch), die auf eine gleichwertige Bewertung abzielen. Die praktizierenden Lehrkräfte sind meist allein für die Vergabe von Abschlussnoten an ihre Schülerinnen und Schülern verantwortlich und somit auch dafür, dass sie die erforderlichen Kenntnisse und Fertigkeiten z. B. in einer Fremdsprache erreicht haben. Dies bedeutet wiederum, dass die unterrichtenden Lehrkräfte im schwedischen System einen vergleichsweise großen Einfluss auf die Bewertung haben (vgl. Nusche et al. 2011).

In der zweiten Fremdsprache stehen nationale Testmaterialien für die Fremdsprachen Deutsch, Französisch und Spanisch zur Verfügung, d. h. die Lehrkräfte können Tests aus einer Prüfungsdatenbank verwenden. Im Gegensatz zu den obligatorischen nationalen Tests im Fach Englisch, sind diese Tests jedoch nur fakultativ. Das Testmaterial wird den Lehrkräften für die Fremdsprachenniveaus Stufe 2, Stufe 3 und Stufe 4 (in etwa A2.1, A2.2 bzw. B1.1 gemäß den Referenzniveaus des GER) auf einer Online-Plattform angeboten, zur Verwendung wird aber nicht aktiv ermutigt. Die Tests folgen generell der traditionellen Einteilung in Hören, Lesen, Sprechen und Schreiben, fokussieren aber in Anlehnung an die Terminologie des GER auf rezeptive Kompetenzen sowie auf mündliche und schriftliche Interaktion und Produktion. Das nationale Testmaterial soll zur Unterstützung der Lehrkräfte bei der Unterrichtsplanung und Entscheidungen darüber dienen, inwiefern die Lernenden am Ende des Kurses die Anforderungen im Lehrplan erfüllen oder nicht. Darüber hinaus zielt das System mit nationalen Testmaterialien in der Fremdsprache darauf ab, die Vergleichbarkeit und die Zuverlässigkeit im Hinblick auf die Bewertung innerhalb der schwedischen Schule zu erhöhen, etwas, was in den letzten Jahren immer häufiger diskutiert wird (vgl. Erickson 2020b).

Zu den jeweiligen Testteilen gehören ausführliche Anweisungen für die Lehrkräfte. Als Unterstützung für die holistische Bewertung schriftlicher Kompetenz werden analytische Beurteilungsfaktoren, die qualitative Aspekte bei der Bewertung von Schülertexten darstellen, sowie mehrere Benchmark-Beispiele bereitgestellt. Dazu kommt, dass die Lehrkräfte Zugang zu weiteren Testmaterialien haben, um die Schülerinnen und Schüler für den Test vorbereiten zu können. Für die Beurteilung der fakultativen Tests gibt es keine externe Kontrolle und diese werden in der Regel von den praktizierenden Lehrkräften

selbst evaluiert (vgl. Håkansson Ramberg 2016). Allerdings wird stark empfohlen, dass die Bewertung dieser Tests in Zusammenarbeit mit Kolleginnen und Kollegen erfolgen sollte (vgl. Skolverket 2021d).

Die Abschlussnote des Kurses für die Fremdsprachen setzt sich aus verschiedenen Aufgaben, Aktivitäten und Klausuren und gegebenenfalls den Ergebnissen der fakultativen Tests zusammen. Die Tatsache, dass alle Arten von Bewertungen in der Regel durch die praktizierenden Lehrkräfte der Schülerinnen und Schüler getroffen werden, lässt Bedenken hinsichtlich der Gerechtigkeit bei der Benotung aufkommen (vgl. Nusche et al. 2011). Des Weiteren könnte die in Schweden auf die jeweiligen Schulen dezentralisierte Herangehensweise zu großen Variationen im Hinblick auf Formen und Methoden für das Beurteilen führen und zudem werden selten detaillierte Informationen darüber, inwiefern Richtlinien zur Qualitätssicherung befolgt wurden, gegeben (ibid.).

2.2.5 Jüngste bildungs- und sprachpolitische Maßnahmen und Diskussionen

Das schwedische Bildungssystem hat in den letzten Jahrzehnten grundlegende Veränderungen erfahren. Zum einen ist die Bildungspolitik in Schweden, wie bereits erwähnt, seit den 90er Jahren durch einen hohen Grad an Dezentralisierung gekennzeichnet, was weitgehend bedeutet, dass die Verantwortung für die allgemeine schulische Ausbildung auf kommunaler Ebene liegt. Dies bedeutet, dass Politiker auf kommunaler Ebene dafür verantwortlich sind, dass die nationalen Ziele für die Schulausbildung und das schwedische Bildungsgesetz befolgt werden. Auf lokaler Ebene wird auch die Verteilung des Budgets beschlossen und daher ist der staatliche ökonomische Einfluss auf die schwedische Schule vergleichsweise gering (Skolverket 2011c). Zum anderen wurde durch die sog. Freischulreform im Jahr 1992 ein System eingeführt, um die Wahlfreiheit von Kindern und Eltern zu erhöhen. Nach diesem System erhalten freie Schulen ebenso wie öffentliche Schulen einen steuerbasierten Beitrag basierend auf der Schüleranzahl. Diese Veränderungen können unterschiedliche Bedingungen für die einzelnen Schulen bedeuten und in weiterer Folge auch für den Fremdsprachenunterricht, z. B. im Hinblick auf Sprachangebot und Anzahl der Schülergruppen. Die dezentralisierte Beschaffenheit des Bildungssystems und die Freischulreform werden deswegen sowohl in Schweden als auch international kritisch erörtert (vgl. Nusche et al. 2011; Molander 2017).

In Schweden scheinen viele Menschen der Meinung zu sein, dass Sprachkompetenz des Englischen ausreichend ist und dass Sprachkenntnisse einer zweiten Fremdsprache nicht von großer Bedeutung sind (vgl. Cabau-Lampa

2007; European Commission 2012a: 70). Auch wenn der Bedarf schwedischer Firmen an Fremdsprachenkenntnissen oft diskutiert wird, werden Sprachkompetenzen in einer zweiten Fremdsprache nicht durchgehend gewährleistet. Wie in vielen anderen europäischen Ländern werden daher auch in Schweden sprachpolitische Diskussionen darüber geführt, wie man das Interesse für das Erlernen moderner Fremdsprachen erhöhen könnte (vgl. Broek & van den Ende 2013). Eine Maßnahme in Schweden ist, dass Schülerinnen und Schüler, die vertiefende Kurse in einer modernen Fremdsprache am Gymnasium belegen, d. h. ihre in der Grundschule gewählte zweite Fremdsprache weiterlernen, zusätzliche Leistungspunkte bekommen können.[36] Diese sog. Meritpunkte geben Abiturienten bessere Chancen, für ein Universitätsstudium zugelassen zu werden (Utbildningsdepartementet 1993). Das Einführen der Meritpunkte im Jahr 2007 (seitdem mehrmals revidiert) hat höchstwahrscheinlich dazu geführt, dass eine höhere Anzahl von Schülerinnen und Schülern ihre in der Grundschule gewählte Fremdsprache im ersten und zweiten Jahr am Gymnasium weiterlernen. Da der darauf aufbauende Kurs, Stufe 5, nur in Ausnahmefällen Meritpunkte gibt, belegen weniger Schülerinnen und Schüler den Kurs (vgl. Utbildningsdepartementet 2018). Die Meritpunkte scheinen einen deutlichen Effekt auf die Anzahl der Schülerinnen und Schüler zu haben, die am Gymnasium ihre in der Grundschule gewählte zweite Fremdsprache weiterlernen, auch wenn dies hauptsächlich in Stadtgebieten wahrgenommen werden kann (vgl. Granfeldt et al. 2021). Trotz Diskussionen (vgl. Gustafsson et al. 2014; Utbildningsdepartementet 2017) sind die Meritpunkte im schwedischen Bildungssystem erhalten geblieben.

In Schweden wird des Weiteren eine Diskussion darüber geführt, ob im schwedischen Bildungssystem eine zweite Fremdsprache als Pflichtfach eingeführt werden sollte, um den Status der zweiten Fremdsprache zu erhöhen. Ausgehend vom Bedarf an Sprachkenntnissen in der schwedischen Gesellschaft wurde 2018 in einem Bericht der schwedischen Schulbehörde vorgeschlagen, dass die Sprachwahl auf freiwilliger Basis in der Grundschule grundsätzlich geändert werden sollte:

> Um das Recht aller Schülerinnen und Schüler auf ihre Muttersprache und zwei weitere Sprachen sowie den Bedarf der Gesellschaft an Sprachkenntnissen besser zu erfüllen, legt die schwedische Schulbehörde Änderungen bei der Wahl einer zweiten

36 Dies gilt ausschließlich für belegte Fremdsprachenstufen in einer modernen Sprache und demzufolge nicht für andere Optionen wie Unterricht in der Muttersprache oder Zeichensprache.

> Fremdsprache in der Grundschule fest, die dazu beitragen können, den Anteil der Schülerinnen und Schüler zu erhöhen, die in der Grundschule moderne Sprachen lernen. Die heutigen Vorgaben ermöglichen es den Schülerinnen und Schülern, das Erlernen moderner Sprachen abzuwählen. Dies schafft eine Ungleichheit, bei der eine große Schülergruppe nicht die gleichen Bedingungen für zwei weitere Fremdsprachen neben der Muttersprache erhält, was keine gleichwertige Ausbildung für alle gewährleistet. (Skolverket 2018a: 22; *eigene Übersetzung, M.H.R.*)[37]

Die Sprachwahl sollte gemäß diesem Vorschlag folgende Optionen enthalten: eine moderne Sprache, Schulunterricht in der Muttersprache für Kinder und Jugendliche mit Migrationshintergrund, Englisch für neu in Schweden angekommene Schülerinnen und Schüler mit geringen Vorkenntnissen oder ohne Kenntnisse in dieser Sprache in der 6. Klasse oder später oder Zeichensprache (Skolverket 2018a). Dabei kann auch die Möglichkeit aller Schülerinnen und Schüler, neben der Muttersprache zwei weitere Fremdsprachen zu erlernen, gestärkt werden. Die schwedische Schulbehörde rechnet damit, dass durch eine verpflichtende zweite Fremdsprache mehr Jugendliche ihre in der Grundschule gewählte Sprache weiterlernen werden, was zu einer erhöhten Sprachkompetenz in Fremdsprachen auf nationaler Ebene führen sollte (ibid.). Eine solche Veränderung würde die Nachfrage nach Lehrkräften für *Moderna språk* beeinflussen und eine Veränderung der Sprachwahl in der Grundschule könnte eventuell den jetzigen Lehrermangel im Bereich der modernen Fremdsprachen verschärfen. Laut einer kürzlich durchgeführten Umfrage scheint heute, im Gegensatz zu früheren Untersuchungen, eine Mehrheit der Lehrkräfte moderner Sprachen in der Grundschule gegenüber einer Reform, die eine zweite Fremdsprache als Pflichtfach einführt, positiv eingestellt zu sein (vgl. Erickson et al. 2018). Darüber hinaus ist eine Revidierung der Lehrpläne für die modernen Sprachen durchgeführt worden. Dadurch sollte sicherstellt werden, dass der Inhalt und die gestellten Anforderungen im heutigen Unterricht erfüllt werden können. Diese revidierte Fassung der Lehrpläne für *Moderna språk* wird für das Schuljahr 2021/2022 am Gymnasium eingeführt (vgl. Skolverket 2021a). Außerdem sollte gemäß dem Bericht untersucht werden, in welcher Beziehung die

37 Im Original: „För att bättre tillgodose alla elevers rätt till sitt modersmål och två ytterligare språk samt samhällets behov av språkkunskaper fastslår Skolverket förändringar i språkvalet i grundskolan som kan bidra till att andelen elever som läser moderna språk i grundskolan ökar. Dagens konstruktion möjliggör för elever att välja bort moderna språk. Detta skapar en ojämlikhet där en stor elevgrupp inte ges samma förutsättningar till två ytterligare språk utöver modersmålet, något som inte gynnar en likvärdig utbildning för alla“ (Skolverket 2018a: 22).

Fremdsprachenstufen des schwedischen Systems zu den Referenzniveaus im GER stehen (Utbildningsdepartementet 2018), eine Aufforderung, die bis heute von Skolverket nicht befolgt wurde.

2.3 Gemeinsamer europäischer Referenzrahmen für Sprachen

Der *Gemeinsame europäische Referenzrahmen für Sprachen: lernen, lehren, beurteilen* (Europarat 2001) wurde zu Beginn der 2000er Jahre nach langjährigen Diskussionen und Kooperationen mehrerer Forscher und Fremdsprachenexperten vom Europarat veröffentlicht und ist seitdem in 40 Sprachen erhältlich. Mit dem GER ist eine neue Plattform für Sprachunterricht und Beurteilung von Sprachkompetenzen in Europa erstanden. Dem GER liegt ein handlungsorientierter Ansatz zu Grunde. Dies bedeutet, dass Sprachverwendende und Sprachlernende einer Sprache als *sozial Handelnde*, die als Mitglieder einer Gesellschaft kommunikative Aufträge in bestimmten Umfeldern und Handlungssituationen bewältigen müssen, gesehen werden. Dies bezieht sich aber nicht nur auf sprachliche Handlungen; erst in einem sozialen Kontext können sie ihre volle Bedeutung erhalten (Europarat 2001: 21 ff.). Der handlungsorientierte Ansatz beachtet somit, dass Lernende eine Vielfalt von Kompetenzen entwickeln und bewältigen müssen, sowohl *allgemeine Kompetenzen*, wie allgemeines Weltwissen oder kognitive Lernfähigkeit, als auch *kommunikative Sprachkompetenzen*. Da die kommunikative Sprachkompetenz gemäß dem GER die linguistische, die soziolinguistische und die pragmatische Kompetenz umfasst, soll ein Lernender daher bei einer Beteiligung an sprachlichen Aktivitäten nicht nur lexikalische, phonologische und syntaktische Kenntnisse und Fertigkeiten einsetzen können, sondern sich auch der Bedeutung gesellschaftlicher Konventionen in der Sprache bewusst sein und die diskursive und funktionale Verwendung sprachlicher Mittel kennen (ibid.).

Der GER ist das Ergebnis einer langjährigen Arbeit im Auftrag des Europarates, aber die Beteiligung des Europarates im Sprachbereich hat bereits wesentlich früher angefangen. Der Europarat wurde 1949 zur Sicherung demokratischer Grundprinzipien gegründet und ist eine europäische Organisation, die der Förderung internationaler Verständigung und Zusammenarbeit dienen soll (vgl. North 2014). Anfang der 70er Jahren wurde das sog. *Threshold Level* (heute das B1-Niveau) definiert, eine Kompetenzstufe, ab der sich ein Fremdsprachenverwender im Land der Zielsprache in der Gesellschaft zurechtfinden kann. Darauf folgten weitere Sprachniveaus wie *Waystage* (heute das A2-Niveau), das als Etappenziel auf dem Weg zum *Threshold Level* festgelegt worden ist, und dies kann als ein erster Versuch der Beschreibung genereller

Referenzniveaus für fremdsprachliche Kompetenz gesehen werden (North 2007: 14 ff.). Eine Konkretisierung dieser Gedanken wurde bei einem Symposium in Rüschlikon in der Schweiz im Jahr 1991 unter dem Titel *Transparency and Coherence in Language Learning in Europe* herausgebildet und die Entwicklung genereller Referenzniveaus zur Förderung der Mobilität von Individuen in Europa über Landesgrenzen hinaus wurde dort beschlossen (vgl. North 2014).

Der Referenzrahmen wurde auch mit der Absicht veröffentlicht, ein transparentes Bezugssystem im Hinblick auf die Vergleichbarkeit unterschiedlicher Bildungssysteme zu erschaffen. Aus diesem Grund wurden genau definierte Deskriptoren, die sog. Kann-Beschreibungen, auf sechs Kompetenzstandards, die zwar ursprünglich als illustrative Beispiele bestimmt waren (Kecker 2014), formuliert. Unterschiedliche Lehrprogramme und Zertifikate im Bereich Fremdsprachenunterricht sollten durch diese Standards oder Sprachkompetenzstufen eine gemeinsame Basis für die Einschätzung fremdsprachlicher Kompetenz erhalten. Zudem sollten diese Kompetenzstandards eine übergreifende Vergleichbarkeit zwischen den Ländern in Europa bezüglich sprachlicher Kompetenz vereinfachen. Dabei ist zu beachten, dass der GER nicht als normierendes Dokument gedacht ist und dementsprechend keine Methoden vorgibt. Dies wird bereits am Anfang des Referenzrahmens von den Autoren klargestellt:

> Wir wollen Praktikern NICHT sagen, was sie tun sollen oder wie sie etwas tun sollen. Wir stellen nur Fragen, wir geben keine Antworten. Es ist nicht die Aufgabe des *Gemeinsamen europäischen Referenzrahmens* festzulegen, welche Ziele die Benutzer anstreben oder welche Methoden sie dabei einsetzten sollten. (Europarat 2001: 8; *Hervorheb. im Original*)

Diese Behauptung kann insofern als widersprüchlich aufgefasst werden, als z. B. ein funktionaler Ansatz im GER gleichzeitig deutlich bevorzugt wird. Die Formulierung scheint eher auf die vielerlei vorhandenen Diskussionen in den Bereichen Spracherwerb und Didaktik hinsichtlich u. a. Lernmethoden und Spracherwerbstheorien hinzudeuten. Der GER ist zudem mit dem ausdrücklichen Ziel veröffentlicht worden, ein Referenzrahmen und eine gemeinsame Plattform für das Sprachlernen in ganz Europa zu sein:

> Der *Gemeinsame europäische Referenzrahmen* stellt eine gemeinsame Basis dar für die Entwicklung von zielsprachlichen Lehrplänen, curricularen Richtlinien, Prüfungen, Lehrwerken usw. in ganz Europa. (Europarat 2001: 14)

Als Referenzpunkt für unterschiedliche Niveaus fremdsprachlicher Kompetenz sollte eine Transparenz zwischen den Ländern im Hinblick auf Lehrpläne, Richtlinien, Sprachkurse und Qualifikationsnachweise erleichtert werden. Der

Referenzrahmen zielt darauf hin, Praktikern und anderen Akteuren im Bildungsbereich einen Rahmen für das Erlernen, Lehren und Beurteilen einer Fremdsprache bereitzustellen:

> Das Dokument ist ein notwendiges Werkzeug für alle, die professionell im Bildungsbereich tätig sind. Didaktikern, Fortbildern, Lehrwerkautoren und Prüfungsexperten dient es bei der Entwicklung von Lehrplänen, Lehrwerken und Sprachprüfungen. (Europarat 2001: 3)

Mit einem gemeinsamen Referenzrahmen für die Anerkennung sprachlicher Kompetenzen sollte eine engere Zusammenarbeit zwischen Sprachpraktikern und Bildungsinstitutionen in Europa geschaffen werden. Nicht zuletzt hat der Referenzrahmen für das Testen und Prüfen von fremdsprachlicher Kompetenz in vielen Bereichen an Bedeutung gewonnen, was auch der Intention der Autoren entspricht: „Außerdem liefert dieses System eine Basis für den Vergleich der zahlreichen Abschlüsse, Kursstufen und Prüfungsniveaus in Europa“ (Europarat 2001: 3).

Der GER ist nicht explizit auf theoretische Modelle, wie z. B. Bachman und Palmers Modell kommunikativer Kompetenz aus dem Jahr 1996 (siehe Kap. 3.2), gegründet, sondern vielmehr auf Kenntnisse und Kompetenzen, die ein Fremdsprachenlerner auf unterschiedlichen Niveaus besitzen sollte. Der Referenzrahmen definiert mündliche und schriftliche produktive, interaktive und rezeptive Fremdsprachenkompetenz und dieses Wissen bezieht sich je nach Stufe auf den privaten oder öffentlichen Bereich. Den Kern des GER bilden demzufolge die Niveaustufen für die sprachliche Kompetenz eines Lernenden. Diese Niveaustufen sind klassisch in die drei Hauptniveaus Grund-, Mittel und Oberstufe untergliedert: „Elementare Sprachverwendung“ (A)[38], „Selbstständige Sprachverwendung“ (B) und „Kompetente Sprachverwendung“ (C), die jeweils in zwei weitere Unterstufen unterteilt sind. Diese insgesamt sechs Referenzniveaus sind in Abb. 2 dargestellt:

---

38 Außerdem finden sich im Begleitband zum GER *Companion Volume with New Descriptors* (seit 2020 in deutscher Übersetzung) zusätzlich neu herausgearbeitete Skalen zum Prä-A1-Niveau (Council of Europe 2020).

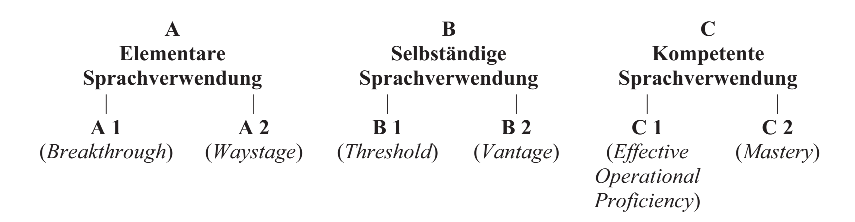


Abb. 2: Die Referenzniveaus des GER (Europarat 2001: 34)

Der GER beschreibt weiterhin die Sprachfähigkeiten, die Fremdsprachenlerner auf diesen sechs Referenzniveaus, von einem elementaren Sprachniveau auf A1 (das niedrigste Referenzniveau) über A2, B1, B2 und C1 bis zum Niveau der kompetenten Sprachverwendung auf C2 (das höchste Referenzniveau) leisten müssen, um kommunikative Ziele auf dem jeweiligen Niveau erfüllen zu können. Die Niveaubeschreibungen finden heutzutage eine weite Verbreitung unter Praktikern im Bildungsbereich, und die Referenzniveaus haben es auch vereinfacht, Sprachleistungen von Fremdsprachenlernenden aus unterschiedlichen Ländern zu vergleichen (vgl. Figueras 2009). Zusätzlich zu einer allgemeinen Beschreibung der jeweiligen Niveaus (die globale Skala), sind auch Teilkompetenzen (die Sub-Skalen), z. B. bezüglich phonologischer und lexikalischer Kompetenz auf den jeweiligen Niveaus, im GER dargestellt. Die globale Beschreibung für das B1-Niveau sieht wie folgt aus:

Tab. 5: *Deskriptoren der B1-Stufe des GER für die Globalskala (Europarat 2001: 35)*

GER-Niveau B1	Kann die Hauptpunkte verstehen, wenn klare Standardsprache verwendet wird und wenn es um vertraute Dinge aus Arbeit, Schule, Freizeit usw. geht. Kann die meisten Situationen bewältigen, denen man auf Reisen im Sprachgebiet begegnet. Kann sich einfach und zusammenhängend über vertraute Themen und persönliche Interessengebiete äußern. Kann über Erfahrungen und Ereignisse berichten, Träume, Hoffnungen und Ziele beschreiben und zu Plänen und Ansichten kurze Begründungen oder Erklärungen geben.

Zu berücksichtigen ist allerdings, dass ein Lernender das Sprachvermögen auf einem elementaren Sprachniveau (A1/A2) häufig schneller erwerben kann als das Sprachvermögen auf einem fortgeschrittenen Niveau (B1–C2). Dies bedeutet, dass die Referenzstufen des GER nicht äquidistant sind (vgl. Quetz &

Vogt 2009; Erickson & Pakula 2017), was demzufolge einen Vergleich anderer Systeme und Modelle mit dem Referenzrahmen erschweren könnte.

Trotz des großen Einflusses des Referenzrahmens als wichtiger Bezugspunkt wurden auch Kritik am GER laut. Es handelt sich dabei z. B. um mangelnde Verankerung in der Forschung (vgl. Harsch 2006), mangelnde Qualität der Leistungsdeskriptoren (vgl. Fulcher 2004; Quetz & Vogt 2009), Probleme in der Terminologie (vgl. Alderson et al. 2006; Harsch 2006) und fehlende empirische Untermauerung der GER-Skalen (vgl. Fulcher 2004; Hulstijn 2007). Darunter fällt auch die allzu große Fokussierung des GER auf Mobilität und Berufsleben, was in erster Linie auf erwachsene Lernende ausgerichtet ist und die Sprachentwicklung und Sprachverwendung von Kindern und Jugendlichen vernachlässigen könnte (Erickson & Pakula 2017). Dazu wird befürchtet, dass der Referenzrahmen in verschiedenen kulturellen Kontexten allzu stark normierend erscheinen könnte (vgl. McNamara 2010). Auch wenn Kritik am GER angeführt worden ist, hatte der Referenzrahmen bislang zweifellos massive Auswirkungen auf das Fremdsprachenerlernen, den Fremdsprachenunterricht und die Bewertung von Fremdsprachenkompetenz in Europa (vgl. Figueras 2009).

2.3.1 Einfluss auf nationale Bildungssysteme

Im Jahr 2008 wurde den europäischen Mitgliedstaaten vom Ministerrat der Europäischen Union empfohlen, den GER in ihren nationalen oder lokalen Bildungssystemen umzusetzen und dabei die Mehrsprachigkeit innerhalb Europas zu fördern (Council of Europe 2008). Dies beinhaltet Bedingungen für eine adäquate Verwendung des GER und dabei sollte der handlungsorientierte und kompetenzbasierte Ansatz im Hinblick auf den Fremdsprachenunterricht innerhalb und zwischen den Mitgliedstaaten berücksichtigt werden. Darüber hinaus wurden nationale, regionale und lokale Bildungsbehörden bei der Umsetzung des GER ermutigt, Bildungsakteure im Fremdsprachenbereich zu koordinieren und bei politischen Entscheidungsprozessen, bei der Curriculumentwicklung, bei der Ausarbeitung von Lehrbüchern, bei der Lehrerausbildung und bei der Bewertung für Vereinheitlichung und Transparenz zu arbeiten. Dabei sollten die zuständigen Behörden auch sicherstellen, dass Prozeduren, die zu offiziellen Sprachleistungsniveaus führten, insbesondere im Hinblick auf den Bezug zu den Referenzniveaus des GER, bei Prüfungen und Bewertungssystemen transparent und zuverlässig vorgenommen werden:

> The CEFR is a reference tool for the development and implementation of coherent and transparent language education policies; when national, regional and local education authorities decide to use it, they are invited to: […]

> 4.5 ensure that all tests, examinations and assessment procedures leading to officially recognised language qualifications take full account of the relevant aspects of language use and language competences as set out in the CEFR, that they are conducted in accordance with internationally recognised principles of good practice and quality management, and that the procedures to relate these tests and examinations to the common reference levels (A1–C2) of the CEFR are carried out in a reliable and transparent manner;
> 4.6 ensure that full information regarding the procedures applied in all tests, examinations and assessment systems leading to officially recognised language qualifications, particularly those used to relate them to the common reference levels (A1–C2) of the CEFR, is published and made freely available and readily accessible to all the interested parties (Council of Europe 2008: 3–4)

Dies bedeutet, dass die Bildungsbehörden der jeweiligen Länder für die Qualität ihrer Prüfungen und Bewertungssysteme verantwortlich sind. Die Behörden müssen folglich vor allem absichern, dass die Prozeduren, die deren Relation zu den Referenzniveaus des GER festlegen, unter validen Verhältnissen durchgeführt werden. Die Zuordnung zu den Referenzniveaus sollte durch Dokumentation unterstützt werden und die Informationen dazu sollten öffentlich zugänglich gemacht werden.

Die Verbindung der Referenzniveaus des GER mit Bildungsstandards oder anderen Dokumenten, die eine Wirkung auf die nationalen Bildungssysteme ausüben, scheint in den europäischen Ländern allerdings nicht immer durch empirische Belege festhalten zu sein (Broek & van den Ende 2013; Bärenfänger 2016). Diese mangelnde Qualitätssicherung könnte aber ein entscheidendes Hindernis sein, wenn der Referenzrahmen auf nationaler Bildungsebene implementiert und verwendet werden soll. Außerdem scheint die Umsetzung des GER auf nationaler Systemebene (in Bildungsgesetzen oder nationalen Lehrplänen) davon abzuhängen, in welchem Grad der GER bei Sprachtests, Lehrmaterialien und in der Lehrerausbildung eingesetzt wird (Broek & van den Ende 2013).

Ein weiteres Problem bei der Implementierung des GER im Hinblick auf Lehrpläne und Bildungsstandards könnte sein, dass die Skalen und Deskriptoren der Referenzniveaus – entgegen der ursprünglichen Intention der Autoren des Referenzrahmens – normierend interpretiert werden (vgl. Quetz & Vogt 2009; Erickson & Pakula 2017). Der Referenzrahmen soll nicht als ein überstaatliches Dokument aufgefasst werden, das Bildungsstandards auf nationaler Ebene reguliert, sondern soll eher als Bezugssystem zur Darstellung von sprachlichen Niveaus dienen. Dies wird auch von einem der Autoren des GER, Brian North, erläutert:

> The main danger with regard to all common frameworks is a simplistic interpretation of them. The key to success is for users to appreciate that a common framework is a descriptive metasystem that is intended as a reference point, not as a tool to be implemented without any further elaboration and adaptation to local circumstances. (North 2007: 10)

Relevant zu beachten ist gemäß North aber auch, dass die Herangehensweise bei einer Implementierung durch weitere und kontinuierliche Verfeinerung dem kulturellen und nationalen Kontext angepasst werden sollte und dass man bei der Implementierung sowohl den Schulkontext als auch das Bildungssystem in den jeweiligen Ländern berücksichtigen muss. Eine Implementierung bedarf demzufolge eines schrittweisen Vorgehens und sollte nicht mit dem ganzen Referenzrahmen beginnen, sondern eher mit der pädagogischen Philosophie und Kultur (North 2014: 111). Inwiefern der GER als Bezugspunkt bei der Implementierung auf nationaler Ebene in den verschiedenen Ländern allzu normierend aufgefasst wurde, sollte aber näher betrachtet werden.

2.3.2 Der GER als Bezugssystem sprachlicher Kompetenz

In den letzten Jahren haben unterschiedliche Bildungsreformen zu einem verstärkten Interesse an externen Standards im Hinblick auf Bildungsstandards und Lehrwerke beigetragen. Den größten Einfluss auf neue nationale Bildungsdokumente und Standards bezüglich Sprachfertigkeitsniveaus in Europa hat der oben erwähnte Referenzrahmen für Sprachen, GER, ausgeübt. In den letzten Jahren hat der GER aber auch das Erstellen nationaler Rahmenwerke und Bildungsstandards in anderen Teilen der Welt beeinflusst (vgl. Schneider et al. 2017). Die Referenzniveaus der Sprachkompetenz (A1–C2) sind dadurch heutzutage weit verbreitet und können zunehmend nicht nur in Europa verstanden und verwendet werden. Durch den GER haben Lernende, Lehrkräfte, Arbeitgeber, Zulassungsbehörden für Sprachstudien sowie andere Interessengruppen ein vergleichbares Instrument und eine Basis für eine genauere Einschätzung der Sprachkompetenz eines Individuums erhalten, was einem der Ziele des GER entspricht:

> Eines der Ziele des Referenzrahmens ist es, allen beteiligten Partnern bei der Beschreibung der Kompetenzniveaus zu helfen, die gemäß den Standards ihrer Tests und Prüfungen erwartet werden. Dies soll den Vergleich zwischen verschiedenen Qualifikationssystemen erleichtern. Zu diesem Zweck sind ein Beschreibungssystem und die Gemeinsamen Referenzniveaus entwickelt worden. (Europarat 2001: 32)

Es wird auch davon ausgegangen, dass ein Vergleich von Sprachkompetenzen durch den Bezug von Prüfungen und Tests auf den GER erleichtert werden

kann. Diese Behauptung wird auch vom Testforscher Michael Kane vertreten. Mit einem Bezugspunkt für die Sprachkompetenz wird gemäß Kane den jeweiligen Ergebnissen, sei es aus einer Sprachprüfung oder einem absolvierten Sprachkurs, eine zusätzliche Bedeutung gegeben: „We can add meaning to the scores by referencing them to […] performance levels, benchmark performance levels, or achievement levels (e.g., as in […] CEFR)" (Kane 2011: 8). Wenn der GER als Bezugssystem verwendet wird, sollten allerdings gewisse Qualitätsanforderungen an Testinstitute und Lehrbuchverlage gestellt werden können.

Validierungsprozesse, die den Bezug zum GER klären sollen, haben vor allem im Bereich nationaler und internationaler Sprachprüfungen an Bedeutung gewonnen. Auch deutschsprachige Institutionen wie das Goethe-Institut, das TestDaf-Institut, das Österreichische Sprachdiplom Deutsch (ÖSD) und die europäischen Sprachzertifikate *TELC* bieten Sprachlernenden ihre Sprachzertifikate gemäß den Referenzniveaus des GER an. Diese hier genannten Prüfungen nehmen alle explizit auf den Referenzrahmen Bezug. Um die Kompetenzniveaus der jeweiligen Prüfenden gleichwertig einschätzen zu können, müssen die Verbindungen einzelner Sprachtests zu den Stufungen des GER jedoch validiert werden. In den letzten Jahren wurden mehrere Studien zur Zuordnung internationaler Sprachtests zu den Referenzniveaus des GER durchgeführt (vgl. Kap. 4.1). Wenn eine solche Qualitätsbestätigung von den einzelnen Instituten intern verfolgt wird, könnte dies aber die Objektivität dieser Validierung in Frage stellen.

Um die Validierungen von Sprachtests zu unterstützen, wurde vom Europarat 2009 die Publikation *Relating language examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR). A manual* (das sog. *Manual*, Council of Europe)[39] herausgegeben. Zudem hat der Sprachtestverband ALTE ein Handbuch, *Manual for language test development and examining* (2011) veröffentlicht. Auch dieses Dokument konzentriert sich auf die Anbindung von Prüfungen an den Referenzrahmen. Beide Dokumente stellen Testanbietern notwendige Methoden zur Qualitätssicherung in einem mehrschrittigen Testverfahren zur Verfügung, das u. a.

39 Eine Pilotversion des Manuals wurde bereits 2003 veröffentlicht, deren methodischer Ansatz allerdings aus vier Phasen besteht. Einigen Studien liegt diese Fassung aus dem Jahr 2003 zugrunde, z. B. O'Sullivan (2008), der in dieser Studie die Validität des Tests *City & Guilds Communicator examination* in Englisch mit dem angepeilten GER-Niveau B2 untersucht hat, und Kecker (2011), die eine Validierungsstudie für den TestDaF (drei TestDaF-Niveaustufen mit Bezug zu den GER-Niveaus B2 und C1) erstellt hat.

Testentwicklung, Testdurchführung und Testbewertung beinhaltet. Der Ansatz im Manual unterscheidet dabei hauptsächlich zwei Prozeduren, die Anbindung des Inhalts, um den adäquaten Verwendungsbereich des Tests zu decken, und die Verknüpfung bestimmter Punktzahlwerte oder anderer Ergebnissen mit den jeweiligen Referenzniveaus.[40]

Des Weiteren bieten europäische und internationale Sprachtestverbände wie EALTA (*European Association for Language Testing and Assessment*) und ALTE (*Association of Language Testers in Europe*) Richtlinien zur Qualitätssicherung für die Bewertung von sprachlichen Kompetenzen (vgl. die Richtlinien der EALTA aus dem Jahr 2006 und die Minimalstandards der ALTE 2007).[41] Diese Testorganisationen können auch externe Qualitätssicherung anbieten. Daher sind z. B. Sprachverbände wie das Goethe-Institut und TELC (die telc gGmbH) seit 1990 bzw. 1995 Mitglieder der ALTE.

Mittlerweile haben sich mehrere standardisierte Sprachtests am GER orientiert und die bekanntesten Zertifikate für Deutsch sind das *Deutsche Sprachdiplom der Kulturministerkonferenz* (abgekürzt DSD), der *Test Deutsch als Fremdsprache* (abgekürzt TestDaF) und die *Goethe-Zertifikate A1–C2* des Goethe-Instituts. Weitere Prüfungen für die deutsche Sprache sind die Sprachzertifikate *TELC* für Deutsch (The European Language Certificates) und die *Deutsche Sprachprüfung für den Hochschulzugang* (DSH). Bisher existiert eine Vielzahl von Studien, die die Beziehung einzelner Sprachtests oder größere Sprachprüfungen, die mit einem Zertifikat verbunden sind, zum Referenzrahmen untersucht haben und dabei ihre Anbindung an die Referenzniveaus des GER vorgeschlagen haben. Viele Institute haben bei der Zuordnung ihrer

---

40 Der systematische Validierungsprozess geschieht nach dem Manual in fünf Phasen, die aufeinander aufbauen: 1) *Familiarisierung* mit dem Referenzrahmen, 2) *Spezifikation* des Tests, 3) *Training* der Standardisierung und Benchmarking, 4) *Standard-Setting* der Leistungen von Lernenden zum GER und 5) empirische *Validierung*, die u. a. durch einen Vergleich zwischen Testergebnissen und Beurteilungen unabhängiger und geschulter GER-Bewertender ablaufen kann (Council of Europe 2009: 113). Generell zeigen vorherige Studien, wie die von O'Sullivan (2008) und Kecker (2011), dass dieser methodische Ansatz zwar gut funktioniert hat, jedoch nicht unproblematisch sei. Beispielsweise konnten nicht alle Aspekte der Deskriptoren bei der Bewertung berücksichtigt werden und einige Aspekte, wie die Aufgabenerfüllung, sind zudem in den GER-Skalen nicht vertreten (vgl. Kecker 2011).

41 Die Organisationen unterscheiden sich u. a. dadurch, dass ALTE sich eher auf Institutionen, nicht auf Individuen, konzentriert, während EALTA ein breiteres Publikum hat und die kollegiale Zusammenarbeit und den Grad von *Assessment Literacy* befürwortet.

Sprachtests zu den Referenzniveaus des GER die methodischen Schritte des vom Europarat herausgegebenen Manuals verwendet (vgl. Council of Europe 2009). Die bereits genannten internationalen Sprachzertifikate für Deutsch der Testinstitutionen Goethe-Institut, TestDaF, Österreichisches Sprachdiplom Deutsch und die telc-Sprachprüfungen wurden alle gemäß unterschiedlichen Qualitätsanforderungen für Sprachprüfungen den GER-Niveaus A1 bis C2 zugeordnet.

Während methodische Ansätze zur Zuordnung unterschiedlicher Sprachtests zu den gemeinsamen Referenzniveaus des GER in den letzten Jahren, vor allem seit der Veröffentlichung des Manuals im Jahr 2009, Aufmerksamkeit erregt haben, ist der Bezug zum Referenzrahmen im Hinblick auf Lehrwerke und Bildungsstandards nicht im gleichen Ausmaß untersucht worden. Seit der Veröffentlichung des Referenzrahmens im Jahr 2001 hat der GER jedoch in mehreren Ländern eine starke Einwirkung auf die Bildungsstandards für die Fremdsprachen gehabt. In vielen Ländern sind heute daher Lehrpläne, Lehrbücher und Sprachtests vorhanden, die sich explizit am GER orientieren.

Der Grad der Umsetzung ist unterschiedlich, aber in vielen europäischen Ländern, wie z. B. Österreich, Frankreich (vgl. Broek & van den Ende 2013) und Finnland (Hildén & Takala 2007; Erickson & Pakula 2017) ist eine starke Berücksichtigung des GER zu sehen. In den letzten Jahren hat der GER Bildungsstandards auf nationaler Ebene auch über die Grenzen Europas hinaus, in Ländern wie z. B. Kanada und Japan (Schneider et al. 2017), beeinflusst. Der Grad der Umsetzung des GER in den verschiedenen Ländern zeigt auch eine Variation im Hinblick darauf, in welchem Ausmaß unterschiedliche Länder den GER in die eigenen Bildungssysteme integriert haben (vgl. Broek & van den Ende 2013; Erickson & Pakula 2017). Grundsätzlich mangelt es für die Anbindung von Lernergebnissen und Bildungsstandards an den GER oft an empirischen Belegen. Dies lässt sich beispielsweise für die Anbindung der Niveaustufen des schwedischen Systems an den GER feststellen, die empirisch als nicht vollständig evaluiert gilt (z. B. Broek & van den Ende 2013; Erickson 2019).

2.4 Umsetzung des GER in Schweden

Schweden hat eine langjährige Tradition im Hinblick auf die Teilnahme an Projekten des Europarats, die das Sprachenlernen und den Fremdsprachenunterricht betreffen, insbesondere in den 70er und 80er Jahren. Spuren dieser Zusammenarbeit, z. B. im Hinblick auf selbständiges Lernen und den funktional-kommunikativen Ansatz im Fremdsprachenunterricht, können

in nationalen Lehrplänen und Lehrwerken erkannt werden. Bereits der im Jahr 1980 eingeführte schwedische Lehrplan, Lgr 80, hatte eine starke kommunikative Prägung und dieser Ansatz ist seitdem in schwedischen Lehrbüchern für den Fremdsprachenunterricht zu spüren (vgl. Andered 2001). Den funktionalen Ansatz gab es in Schweden demzufolge bereits vor der Entstehung des GER.

Der europäische Referenzrahmen hatte aber einen deutlichen Einfluss auf die Gestaltung der schwedischen Bildungsstandards bezüglich Fremdsprachen. Der Bezug zum GER ist zudem durch die jüngsten Reformen in zunehmendem Grad expliziter geworden und der Fokus auf kommunikative Kompetenz hat sich verstärkt. Bereits die schwedischen Bildungsstandards für die Fremdsprachen aus dem Jahr 2000 zeigen eine Beziehung zum bald darauf erschienen Referenzrahmen[42], beispielsweise im Hinblick auf die Terminologie und einen verstärkten Fokus auf die interaktionale Kompetenz (Skolverket 2012). Darüber hinaus wurde, wie bereits erwähnt, ein gemeinsames System für die Progression der Fremdsprachen in der Grund- und Gymnasialschule mit Bezug auf den Referenzrahmen eingeführt.

Die schwedische Fassung des GER wurde im Jahre 2009, u. a. in Vorbereitung auf die Reform der neuen Lehrpläne für die modernen Fremdsprachen im Jahr 2011, veröffentlicht (Erickson & Pakula 2017). Der GER ist in Schweden jedoch nicht in einem rechtlichen bindenden Dokument, wie dem Bildungsgesetz, den nationalen Lehrplänen oder den Lehrplänen für Fremdsprachen umgesetzt. Auch wenn der GER demzufolge in den heutigen nationalen Lehrplänen für die modernen Fremdsprachen nicht explizit erwähnt wird, kann der Einfluss des Referenzrahmens in den nationalen Dokumenten dennoch erkannt werden. Es handelt sich z. B. um einen verstärkten handlungsorientierten Ansatz zum Spracherwerb, um Texttypen und Kontexte des Sprachgebrauchs und um Terminologie. Erst im Kommentarmaterial zum Lehrplan für *Moderna språk* aus dem Jahr 2011 wird auf den Einfluss des Referenzrahmens auf die schwedischen Lehrpläne eingegangen (vgl. Skolverket 2011b). Diese explizite Erwähnung des GER in den schwedischen nationalen Bildungsstandards wurde vergleichsweise relativ spät eingeführt (Skolverket 2012), was die Umsetzung des Referenzrahmens in Schweden womöglich verzögert hat.

In Schweden findet der GER jedoch immer noch nicht in allen Bereichen der Sprachausbildung Berücksichtigung und die Variabilität scheint dabei

42 Basierend auf einem Vorgänger des GER sowie auf eine Pilotversion des GER (vgl. Kap. 2.4.2).

groß zu sein, vor allem im Hinblick darauf, inwieweit Lehrkräfte das Dokument überhaupt kennen oder verwenden. Es scheint aber auch andere Gründe für die verzögerte Umsetzung des GER zu geben. Ein Grund könnte die starke Verantwortung der schwedischen Lehrkräfte für den eigenen Sprachunterricht und die Beurteilung von Schülerleistungen sein. Dazu ist die Arbeitsbelastung der Lehrkräfte während des Schuljahrs häufig sehr hoch. Die Lehrkräfte haben daher womöglich keine Zeit, sich dem Referenzrahmen zu widmen, und müssen dies auch nicht tun, um Unterricht und Bewertung durchzuführen. Fortbildungen für Sprachlehrkräfte in diesem Bereich kommen außerdem selten vor. Des Weiteren verfügen Lehrbücher und Lehrmaterialien schwedischer Verlage für die Schule selten über einen Hinweis auf den Referenzrahmen, was aber erstaunlich ist, da der GER gemäß den Lehrplänen aus dem Jahr 2011 einen Bezugsrahmen für die Standards der modernen Sprachen in Schweden bildet.

Andere Länder, wie z. B. Finnland, haben deutlicher Bezug auf den GER genommen (vgl. Hildén & Takala 2007), sodass der Referenzrahmen mittlerweile ein etabliertes Dokument innerhalb des finnischen Sprachunterrichts ist. In Finnland sind die Referenzniveaus des GER in die Bildungsstandards integriert worden und die Zuordnung von Sprachprüfungen zum GER ist zudem durch empirische Belege dokumentiert worden (vgl. Erickson & Pakula 2017). Auch in Schweden werden Diskussionen darüber geführt, den Bezug der Fremdsprachenstufen im schwedischen System zum GER zu evaluieren und zu verdeutlichen. Jedoch ist es noch zu früh, Aussagen darüber zu treffen, inwieweit Schweden diesbezüglich dem Beispiel Finnlands folgt.

2.4.1 Schwedische Bildungsstandards für die Fremdsprachen und deren Bezug zum GER

Die aktuellen schwedischen Bildungsstandards für Fremdsprachen an der Grundschule und am Gymnasium wurden im Jahr 2011 eingeführt.[43] Diese Reform bedeutete eine Konkretisierung der bereits vorhandenen Lehrpläne, beinhaltete aber auch einen Übergang von der seit 1994 existierenden viergradigen Skala (IG–MVG) zu einer sechsgradigen Bewertungsskala, wonach die Noten E–A als bestandene Noten und die Note F als ungenügend gelten. Schwedische Lehrpläne sind seit der Reform 1994 zielorientiert und setzen auf ein kriterienbasiertes Bewertungssystem (vgl. Gustafsson & Erickson 2013),

43 Überarbeitete Versionen der Lehrpläne für die Schulfächer Englisch sowie *Moderna språk* gelten ab 1. Juli 2021 für Gymnasium und Erwachsenenbildung und ab 1. Juli 2022 für die Grundschule.

was häufig bedeutet, dass konkrete Kompetenzen definiert werden, die für jede einzelne Notenstufe erreicht werden müssen. Schulische Leistungen werden somit gegen diese inhaltlich formulierte Ziele und Kriterien, die Schülerinnen und Schüler am Ende des Kurses erreicht haben sollten, geprüft. Dieses System ersetzte das alte seit Anfang der 50er Jahren existierende normorientierte Benotungssystem.[44] Die jeweiligen Kurse am Gymnasium werden, wie auch die Fächer in der Grundschule, nach der sechsgradigen Skala F–A benotet.

Wie schon in früheren Lehrplänen seit den 80er Jahren zum Ausdruck gekommen ist, wird der Fokus im schwedischen Fremdsprachenunterricht auf die kommunikative Sprachkompetenz gelegt (vgl. Erickson 2019). Dies wird bereits in der Einleitung der Lehrpläne für Fremdsprachen am Gymnasium deutlich:

> Der Unterricht im Fach *Moderna språk* sollte darauf abzielen, dass die Schülerinnen und Schüler ihre Kenntnisse in der Zielsprache und dem allgemeinen Weltwissen sowie ein Vertrauen in ihre Fähigkeit, die Sprache in verschiedenen Situationen und für verschiedene Zwecke verwenden zu können, entwickeln. Den Schülerinnen und Schülern sollte die Möglichkeit gegeben werden, durch Sprachverwendung in funktionalen und sinnvollen Kontexten eine allumfassende kommunikative Fähigkeit zu entwickeln. (Skolverket 2011a; *eigene Übersetzung, M.H.R*)[45]

In den schwedischen Bildungsstandards für Fremdsprachen werden zentrale Inhalte definiert und zentrale Bildungsziele, die die Schülerinnen und Schüler bis zum Ende jeder Lerneinheit (*kurs*) erworben haben sollten, benannt. Diese umfassen die Teilbereiche *Rezeption*, d. h. die Fähigkeit, gesprochene und geschriebene Sprache zu verstehen, sowie *Produktion und Interaktion*, d. h. einerseits die Fähigkeit, sich mündlich und schriftlich in einer Kommunikation, die auf Sendern basiert ist (die Produktion), und andererseits in einer interaktiven Kommunikation, die dialogisch orientiert ist (die Interaktion) adäquat und angemessen auszudrücken. Jene Einteilung in rezeptive Fähigkeiten bzw.

---

44 In einem normorientierten Bewertungssystem werden schulische Leistungen von Individuen oder Gruppen mit einer Bezugsnorm, z. B. mit einer anderen Bezugsgruppe oder der Gesamtpopulation, verglichen. Hierbei soll nicht festgelegt werden, *was* die Lernenden am Ende vom Kurs können, sondern eher in welchem Verhältnis diese schulischen Leistungen *zu anderen Leistungen* stehen.

45 Im Original: „Undervisningen i ämnet moderna språk ska syfta till att eleverna utvecklar kunskaper i målspråket och omvärldskunskaper samt tilltro till sin förmåga att använda språket i olika situationer och för skilda syften. Eleverna ska ges möjlighet att, genom språkanvändning i funktionella och meningsfulla sammanhang, utveckla en allsidig kommunikativ förmåga" (Skolverket 2011a).

mündliche und schriftliche Interaktion und Produktion statt der traditionellen Einteilung in Hören, Lesen, Schreiben und Sprechen ist ebenfalls im Einklang mit dem Referenzrahmen umgesetzt worden (Erickson & Pakula 2017). Des Weiteren umfassen die zentralen Inhalte auch Kontexte und Texttypen für Sprachverwendung, die ihre Entsprechung im GER finden (Skolverket 2011b). Durch den Einfluss des GER hat sich der funktionale Schwerpunkt verstärkt und die schwedischen Lehrpläne für die modernen Sprachen weisen heute noch stärker als zuvor auf einen praxisorientierten kommunikativ-funktionalen Ansatz für den Fremdsprachenunterricht hin (ibid.).

2.4.2 Zuordnung der schwedischen Fremdsprachenstufen zu den GER-Niveaus

Seit der Bildungsreform im Jahr 2000 hat Schweden ein gemeinsames System für die Progression der Fremdsprachen in sowohl der Grund- als auch der Gymnasialschule. Im den Lehrplänen zugehörigen Kommentarmaterial aus dem Jahr 2011 wird erwähnt, dass sich die Sprachstufen des schwedischen Bildungssystems an den Referenzniveaus des GER orientieren:

> Ein wichtiger Bestandteil bei der Ausarbeitung des neuen Lehrplans für *Moderna språk* war es, dass er, wie vorher, Teil eines mit der Gymnasialschule gemeinsamen Systems sein sollte, das generelle und aufeinander aufbauende Sprachniveaus, sog. „steg" (Stufen) enthält. Ausgangspunkt dieses Systems ist der vom Europarat herausgegebene „Gemeinsame Europäische Referenzrahmen für Sprachen: lernen, lehren, beurteilen" (GER). Dies ist ein anerkanntes europäisches System mit generellen Sprachniveaus. (Skolverket 2011b: 6; *eigene Übersetzung, M.H.R.*) [46]

Die Absicht war folglich, dass die GER-Niveaus als Ausgangspunkt für das schwedische Bildungssystem dienen sollten. In Schweden haben ein Vorgänger des GER (Holec et al. 1996) sowie eine Pilotversion des GER eine Basis für das Erstellen der neuen Lehrpläne für den Fremdsprachenunterricht aus dem Jahr 2000 geboten, wobei die Progression bezüglich der Fremdsprachen in sieben verschiedene Niveaustufen eingeteilt wurde. Diese neue Einteilung in sieben Niveaus bedeutete eine Annäherung an die sechs Referenzniveaus des

46 Im Original: „Ett viktigt inslag vid utarbetandet av den nya kursplanen i moderna språk har varit att den, liksom tidigare, ska ingå i ett med gymnasieskolan gemensamt system med generella och påbyggbara språknivåer, så kallade steg. Utgångspunkten för detta system är Europarådets „Gemensam europeisk referensram för språk, lärande, undervisning och bedömning" (GERS). Detta är ett vedertaget europeiskt system med generella språknivåer."

Frameworks und später auch des GER, geschah allerdings vor dem offiziellen Erscheinen des GER. Da das schwedische System aus sieben Fremdsprachenniveaus besteht, kann es jedoch nicht direkt auf die sechs Niveaustufen des GER übertragen werden. Hierzu ist zunächst auch zu bemerken, dass bestimmte GER-Niveaus mit einer ausreichenden Note E der jeweiligen Sprachstufen vergleichbar sind. Die schwedischen Sprachstufen reichen aber nicht bis zu den höheren C1- und C2-Niveaus im Referenzrahmen und sind daher nicht gleich umfassend wie der GER. Dies hat u. a. damit zu tun, dass die schwedischen Fremdsprachenstufen dem ganzen Schulsystem angepasst werden sollten und dass diese Niveaus gleichzeitig auch mit den Stufen für das Fach Englisch im schwedischen System zusammenpassen sollten (vgl. Erickson & Pakula 2017). Aus diesen Gründen sind die GER-Niveaus in Unterstufen unterteilt, wenn diese mit den schwedischen Fremdsprachenstufen verglichen werden sollen.

Der Bezug zum GER kommt allerdings in den Bildungsstandards nicht eindeutig zum Ausdruck. Auch wenn die Aufgabenstellungen des nationalen Prüfungsmaterials betrachtet werden, lässt sich feststellen, dass keine Hinweise zu den Referenzniveaus des GER vorhanden sind. Die Anbindung an den Referenzrahmen wird aber in anderen herausgegebenen Zusatzmaterialen beschrieben. Im Kommentarmaterial der schwedischen Lehrpläne für *Moderna språk* etwa werden die Sprachstufen des schwedischen Systems mit den Referenzniveaus im GER verlinkt (vgl. Skolverket 2011b). Die tentative Zuordnung der GER-Niveaus (A1–B2) zu den Fremdsprachenstufen (1–7) des schwedischen Bildungssystems ist in Tab. 6 unten ersichtlich:

Tab. 6: *Niveaustufenüberblick der Relation zwischen schwedischen Fremdsprachenstufen und den Referenzniveaus des GER (nach Skolverket 2011b: 7)*

GER-Niveau	*A1.1*	*A1.2*	*A2.1*	*A2.2*	*B1.1*	*B1.2*	*B2.1*	*B2.2*
Grundschule		Wahlfach (Klasse 8–9)	Sprachwahl (Klasse 6–9)					
Sprachstufen am Gymnasium		1	2	3	4	5	6	7

Wenn in der 6. Klasse der Grundschule mit einer modernen Fremdsprache begonnen wird, schließen die Schülerinnen und Schüler nach vier Jahren den Unterricht in der 9. Klasse mit der zweiten Niveaustufe ab, was mindestens einem erreichten A2.1-Niveau gemäß dem GER entsprechen sollte. Um die jeweiligen GER-Niveaus zu erreichen, müssen die Lernenden mindestens eine ausreichende Note E in dem entsprechenden schwedischen Fremdsprachenniveau

erhalten haben (Skolverket 2011b; Oscarson 2015). Dies bedeutet demzufolge, dass eine ausreichende Note E in der fünften Niveaustufe des schwedischen Bildungssystems äquivalent zu einem erreichten B1-Niveau des GER sein sollte.

Auch wenn die schwedischen Bildungsstandards für die Fremdsprachen sich bereits seit dem Jahr 2000 auf die Referenzniveaus des GER beziehen, sind nur wenige Validierungsstudien im Hinblick auf das Verhältnis der Sprachniveaus der schwedischen Schülerinnen und Schüler zu den Referenzniveaus des GER durchgeführt worden. Allerdings sind tentative Übereinstimmungsstudien zwischen den Fremdsprachenstufen des schwedischen Bildungssystems und den Referenzniveaus des GER im Auftrag der schwedischen Schulbehörde als interne Berichte durchgeführt worden. Insgesamt drei textuelle Analysen von den Forschenden Mats Oscarson[47], Raili Hildén und Lena Börjesson haben die Zuordnung der schwedischen Fremdsprachenstufen zu den Referenzniveaus des GER, die im Kommentarmaterial ersichtlich ist, auf textueller Ebene untersucht. Folglich wurden in den Analysen lediglich die Formulierungen der Bildungsstandards mit den entsprechenden Skalen des GER verglichen und sie bauen daher nicht auf empirischen Testergebnissen auf, die auf die jeweiligen GER-Niveaus bezogen werden könnten. Die Studien sind zu unterschiedlichen Zeitpunkten im Auftrag der schwedischen Schulbehörde als interne Berichte für die Behörde verfasst worden (vgl. Erickson & Pakula 2017). Diese internen Berichte sind von der schwedischen Schulbehörde nicht veröffentlicht worden

47 Oscarson (2015) beschreibt in seiner späteren Publikation, basierend auf seinem internen Bericht aus dem Jahr 2002, das Prozedere, wie die sieben Fremdsprachenstufen der schwedischen Lehrpläne im Hinblick auf Inhalt und Struktur den sechs Referenzniveaus zugeordnet wurden. Die im Auftrag von Skolverket 2001–2002 durchgeführten Analysen seien schrittweise durchgeführt worden. Um einen genaueren Vergleich zu erstellen, wurde neben der übergreifenden Zielsetzung („*mål att uppnå*") auch die Wissensanforderungen („*kunskapskrav*") für die niedrigste Bestehensgrenze, d. h. eine ausreichende Note E jeder einzelnen Fremdsprachenstufe des schwedischen Systems, in die Analyse miteinbezogen. Durch diese Analysen konnte bestätigt werden, dass die sieben Fremdsprachenstufen der schwedischen Schule im Hinblick auf eine ausreichende Note E sich vom A1-Niveau bis zu einem B2-Niveau erstrecken, aber auch, dass rezeptive Fertigkeiten auf einem höheren Kompetenzniveau als produktive eingeschätzt werden konnten. Für die schriftliche Produktion fanden sich die Fremdsprachenstufen gemäß den textuellen Analysen auf folgenden Referenzniveaus: „*Steg 1*": A1–A2; ‚"*Steg 2*": A2; „*Steg 3*": A2–B1, „*Steg 4*": B1 (niedrig); „*Steg 5*": B1; „*Steg 6*": B2 (niedrig); „*Steg 7*": B2 (Oscarson 2015: 141), was somit weitgehend der abschließenden Zuordnung der Fremdsprachenstufen zu den GER-Niveaus entspricht (vgl. Skolverket 2011b).

und nur die Studie von Hildén (2008) und eine spätere Publikation von Oscarson (2015) waren für die vorliegende Studie zu erhalten. Die Analysen des internen Berichts von Börjesson aus dem Jahr 2009, werden dagegen in keiner später erschienenen Publikation beschrieben.[48]

Die durch die textuellen Analysen eher tentative Anbindung der schwedischen Lehrpläne an den GER wird als problematisch angesehen, da diese Beziehung nicht genügend validiert oder empirisch belegt worden ist. Der Mangel an empirischen Befunden wird u. a. in einem Bericht der europäischen Kommission aus dem Jahr 2013 erläutert:

> in Schweden […] wurde auf das *Fehlen empirischer Beweise* hingewiesen, da keine Forschungsstudien durchgeführt wurden, um den Zusammenhang zwischen dem GER und den Dokumenten und Prüfungen nachzuweisen, die sich am GER orientieren. (Broek & van den Ende 2013: 42, *Hervorheb. im Original*)

Da der Fokus der bisherigen textuellen Studien vor allem auf die Terminologie und die Formulierungen in den Bildungsstandards gelegt wurde und wie diese an die Terminologie und die Formulierungen in den Deskriptoren des GER anknüpfen, werden in der Forschung weitere empirische Studien empfohlen (vgl. Erickson & Pakula 2017).

48 Die Tatsache, dass die Berichte der schwedischen Schulbehörde zur Anbindung der Fremdsprachenstufen an den GER von Mats Oscarson und Lena Börjesson nicht frei zugänglich sind, steht im Widerspruch zu den Empfehlungen der Europäischen Kommission für einen freien Zugang zu Dokumenten im Hinblick auf den Zuordnungsprozess (Council of Europe 2008), die auf ein transparentes und valides Prozedere zeigen sollten (vgl. Kap. 2.3).

3. Konzeptioneller Rahmen

Die vorliegende Arbeit untersucht ausgehend vom schwedischen Schulkontext Bewertungen fremdsprachlicher Kompetenz im Hinblick auf die Sprachfähigkeit in schriftlicher Produktion und Interaktion. Durch Bewertungen von Lernerleistungen in einem Sprachtest und die Interpretation jener Testergebnisse können Aussagen über die kommunikative Sprachfähigkeit eines Lernenden erörtert werden und somit soll vorausgesagt werden, wie der Lernende spezifische Alltagssituationen in der Fremdsprache bewältigen kann. Die kommunikative Sprachfähigkeit durch einen Test zu bewerten, der eine authentische Situation darstellen soll, kann allerdings sehr komplex sein. Zum einen ist es schwierig, authentische Aufgaben für eine Testsituation zu erstellen. Des Weiteren entsteht zum anderen ein Risiko, dass die Interpretationen und Schlussfolgerungen, die wir über die Sprachfähigkeit eines Individuums auf Basis des Testergebnisses ziehen, an Wert verlieren, da der Lernende nur in diesem bestimmten Kontext einen Nachweis dafür gebracht hat. Zudem können Faktoren oder Merkmale, die als irrelevant zu betrachten sind, die Bewertung beeinflussen. Die Gefahr, dass andere Faktoren oder Merkmale eine Bewertung beeinflussen, ist häufig insbesondere im Hinblick auf die freie Produktion ernst zu nehmen. Daher ist von Gewicht, dass wir die Interpretation und die Verwendung der Testergebnisse sowie in weiterer Folge auch deren Konsequenzen auch rechtfertigen können (*Validität*). Es ist aber auch wichtig, dass die Aussagen und Entscheidungen, die über die Sprachfähigkeit eines Lernenden getroffen werden, zuverlässig und reliabel sind. Wie Aspekte der Validität – insbesondere im Hinblick auf eine Bewertung fremdsprachlicher Kompetenz aus der Perspektive der Bewertenden verstanden werden können, soll zunächst diskutiert werden.

Eine Herausforderung im Bereich des Fremdsprachentestens ist allerdings, *was* bewertet werden soll und wie dies zu definieren ist. Um ein besseres Verständnis davon zu erhalten, wie Kenntnisse in einer Fremdsprache (L2), die in einer großen Anzahl von Bereichen und Kontexten erlangt werden können, zu verstehen sind, wurden unterschiedliche theoretische Modelle vorgeschlagen. Diese Modelle enthalten Komponenten, die die menschliche kommunikative Sprachfähigkeit gestalten und definieren sollen, und sind oft als Basis für die Konstruktion eines Tests und dessen Bewertung, Interpretation und Verwendung gedacht. Man kann sich allerdings der Kompetenz in einer Fremdsprache auch auf andere Weisen als durch theoretische Modelle unterschiedlicher

Komponenten annähern. Eine Vorgehensweise ist beispielsweise, verschiedene Standards dafür aufzustellen, welche Sprachfähigkeiten Fremdsprachenlernende auf einem bestimmten Niveau erreicht haben sollten. Nachdem bereits in den vorherigen Kapiteln der handlungsorientierte Ansatz des GER aufgegriffen wurde, soll in diesem Kapitel auf die Definition und den Entwicklungsverlauf des Begriffes *kommunikative Kompetenz* eingegangen werden. Des Weiteren werden grundlegende Konzepte und Kompetenzmodelle einer solchen Kompetenz sowie deren Entwicklung und Einfluss im Bereich Fremdsprachenunterricht erläutert (Kap. 3.1).

Welche Indikatoren weisen auf Qualität bei einer Bewertung von Sprachkompetenzen hin? Diese Frage lässt sich in verschiedenen Paradigmen zum Teil unterschiedlich beantworten. Wenn die Qualität einer Bewertung bestimmt werden soll, beinhalten die Untersuchungen häufig Begriffe wie Validität (Cureton 1951; Cronbach 1971), Testnützlichkeit (Bachman & Palmer 1996), Reliabilität (vgl. Johansson 2015; Tengberg et al. 2017) und Validitätsargumente (Chapelle et al. 2008; Kane 2013). Die Validität und die damit verbundenen Konzepte gelten heute als der wichtigste Ansatz im Hinblick auf Untersuchungen zur Qualität bei einer Bewertung. Traditionell beschäftigt man sich in der Auseinandersetzung mit Validität mit der Frage, inwiefern ein Test das misst, was er messen sollte. Leitende Forscher beschreiben jedoch diese Definition als allzu begrenzt und berücksichtigen in ihren Definitionen auch die Interpretationen und die Verwendung der Testergebnisse (z. B. Messick 1989a; Kane 2001; Moss et al. 2006). Zunächst wird auf für diese Studie relevante Konzepte, Theorien und Rahmenmodelle im Hinblick auf die Validität bei der Interpretation und Verwendung eines Tests und dessen Testergebnissen eingegangen (Kap. 3.2). Als ein zentraler Qualitätsindikator bei einer Bewertung wird häufig auch die Reliabilität angesehen. Wenn Testergebnisse interpretiert und verwendet werden sollen, ist von Bedeutung, einen Indikator ihrer Reliabilität zu haben. Die Reliabilität wird in der Forschung gelegentlich als ein eigener Qualitätsindikator verstanden, ist aber häufig nach einer einheitlichen Definition von Validität im Validitätsbegriff einbegriffen. In diesem Kapitel wird jedoch die Reliabilität in einem eigenen Abschnitt dargestellt, um auf Urteilstendenzen bei der Bewertung von Textproduktionen und verschiedene Kategorien hinsichtlich der Reliabilität eingehen zu können (Kap. 3.3).

3.1 Kompetenz und Kompetenzmodelle

Die Entwicklung unterschiedlicher Kompetenzen ist in schulischer Bildung ein wichtiges Ziel. Inwiefern Lernende diese Kompetenzen am Ende des

Schuljahres erreicht haben oder nicht, wird gegen die konkretisierten Anforderungen in den jeweiligen Lehrplänen evaluiert. Das Interesse für Kompetenzmessung hat sich in den letzten Jahrzehnten erhöht, nicht zuletzt durch internationale Bildungsvergleichsstudien wie PISA und TIMMS. Die Ergebnisse dieser Kompetenzmessungen bilden außerdem auf nationaler Ebene eine Grundlage für bildungspolitische Diskussionen. Erforderliche Kompetenzen zu definieren sowie theoretisch und empirisch gegründete Kompetenzmodelle zu entwickeln, wird nach wie vor als eine Herausforderung angesehen. Um Kompetenzen unter Lernenden in verschiedenen Bildungssystemen vergleichen zu können, sind zudem bisher kaum gemeinsame Kompetenzmodellierungen zwischen Ländern herausgearbeitet worden (Klieme & Leutner 2006), nach welchen sich z. B. die nationalen Bildungsstandards orientieren könnten. Einige wenige Arbeiten in diese Richtung sind zu finden, wie beispielsweise der europäische Referenzrahmen (Europarat 2001) für das Erlernen einer Fremdsprache, der zunehmend für Kompetenzmessung verwendet wird. Die Definition von Kompetenzen und die Skizzierung von Kompetenzmodellierungen bleibt jedoch weiterhin auf sowohl nationaler als auch internationaler Ebene eine zentrale Frage für die Forschung.

Das vorliegende Kapitel wird kurz den Schlüsselbegriff Kompetenz sowie den Entwicklungsverlauf kommunikativer Sprachkompetenz erörtern (Kap. 3.1). In Bezug darauf widmet sich das nächste Kapitel Modellierungen kommunikativer Sprachkompetenz und deren Einfluss im Bereich des Fremdsprachentestens. Genauer beschrieben wird hier die Kompetenzmodellierung des GER, da sich die schwedischen Bildungsstandards am GER orientieren (Kap. 3.2).

3.1.1 Definition und Entwicklung kommunikativer Kompetenz

Eine allgemeine und in der Bildungsforschung häufig zitierte Definition des Begriffs *Kompetenz* stellt der Erziehungswissenschaftler Weinert (2001) auf. Unter „Kompetenzen" versteht Weinert:

> […] die bei Individuen verfügbaren oder durch sie erlernbaren kognitiven Fähigkeiten und Fertigkeiten, um bestimmte Probleme zu lösen, sowie die damit verbundenen motivationalen, volitionalen und sozialen Bereitschaften und Fähigkeiten, um die Problemlösungen in variablen Situationen erfolgreich und verantwortungsvoll nutzen zu können (Weinert 2001: 27–28)

In Anlehnung an frühere Forschung definiert er hierbei den Kompetenzbegriff als deklaratives Wissen (Wissen über Sachverhalte) und prozedurale Fertigkeiten (Wissen, wie man etwas tut), eine bestimmte Aufgabe zu lösen. Weinert erweitert zudem die Definition, indem er die kognitiven Fähigkeiten

und Fertigkeiten, d. h. das Wissen und Können, auch mit der Motivation, dem Willensvermögen sowie der sozialen Bereitschaft eines Individuums kombiniert und wie adäquat diese Faktoren in problemlösenden Situationen eingesetzt werden. Für das Bildungswesen ist dabei wichtig, dass die Kompetenzen im Wesentlichen durch Erfahrung und Erlernen erworben werden können (vgl. Klieme & Leutner 2006).

Der Kompetenzbegriff ist, obwohl schon vor Jahrzehnten bekannt gemacht, immer noch Gegenstand fortdauernder Diskussionen. Der Begriff wurde von Noam Chomsky (1965) in die Sprachwissenschaft eingeführt. Chomsky knüpft an die Unterscheidung de Saussures von *Language* und *Parole* an und stellt die Aufteilung in Kompetenz (*competence*), die er als die allgemeine Sprachfähigkeit versteht, und Performanz (*performance*), definiert als die Verwendung der Sprache in authentischen Situationen, dar. Im Mittelpunkt von Chomskys Kompetenz steht die Beherrschung grammatischer Regeln, unabhängig von pragmatischen, soziolinguistischen oder semantischen Einflüssen.

Anfang der 70er Jahre hat Dell Hymes das Konzept von *Kompetenz* u. a. mit soziolinguistischen Faktoren erweitert und dabei den Begriff *kommunikative Kompetenz* konzipiert (1972). Er versteht darunter nicht nur grammatisches Wissen, sondern berücksichtigt auch die Fähigkeit, Sprache kontextbezogen und soziolinguistisch adäquat zu verwenden. Hymes Definition von kommunikativer Kompetenz hat seitdem einen großen Einfluss auf die Methodik für das Erlernen und Testen von Fremdsprachen ausgeübt und markiert in vielerlei Hinsicht einen Paradigmenwechsel, der häufig auch als die *kommunikative Wende* bezeichnet wird. Dieser Paradigmenwechsel in der Fremdsprachendidaktik verlangte Modelle fremdsprachlicher Kompetenz, die Sprachverwendung in den Vordergrund stellen und als Basis für das Fremdsprachentesten funktionieren sollten. Seitdem entstand eine Reihe von Modellen und Definitionen fremdsprachlicher Kommunikationsfähigkeit.

Mehrere spätere Modelle der Sprachkompetenz bauen auf die Definition kommunikativer Kompetenz von Hymes' auf (vgl. Canale & Swain 1980; Bachman & Palmer 1996). Auch der Kompetenzbegriff im europäischen Referenzrahmen (2001) greift auf Hymes Definition von kommunikativer Kompetenz zurück und basiert zugleich deutlich auf einem handlungsorientierten Ansatz:

> Kompetenzen sind die Summe des (deklarativen) Wissens, der (prozeduralen) Fertigkeiten und der persönlichkeitsbezogenen Kompetenzen und allgemeinen kognitiven Fähigkeiten, die es einem Menschen erlauben, Handlungen auszuführen. (Europarat 2001: 21)

Hier steht die Sprachverwendung im Vordergrund: Sprachlernende werden gemäß dem GER als soziale Akteure angesehen, die innerhalb von „spezifischen Umgebungen und Handlungsfeldern kommunikative Aufgaben" (Europarat 2001: 21) bestehen müssen. Um diese kommunikativen Aufgaben in bestimmten Lebensbereichen (*Domänen*) bewältigen zu können, benötigt der Sprachlernende sowohl *allgemeine Kompetenzen* als auch *kommunikative Sprachkompetenzen* (ibid.). Im GER wird die allgemeine Kompetenz vier Wissens- und Könnenskategorien zugeordnet. Auf einer ersten Ebene werden hier deklaratives Wissen (*savoir*), Fertigkeiten und prozedurales Wissen (*savoir-faire*), persönlichkeitsbezogene Kompetenz (*savoir-être*) sowie Lernfähigkeit (*savoir-appendre*) unterschieden (vgl. Europarat 2001: Kap. 5.1). Diese enthalten jedoch im GER, anders als die Teile der kommunikativen Sprachkompetenz, keine weiteren Deskriptoren oder Kompetenzniveaus. Die kommunikative Sprachkompetenz wird im nächsten Abschnitt näher erörtert.

3.1.2 Sprachkompetenzmodelle und die Orientierung an externen Sprachstandards

In der Bildungsforschung wird in der Regel zwischen Kompetenzmodell als *Kompetenzstrukturmodell* und Kompetenzmodell als *Kompetenzniveaumodell* differenziert (vgl. Klieme & Leutner 2006). Kompetenzstrukturmodelle beschreiben einerseits, welche und wie viele verschiedene Kompetenzdimensionen von Sprachlernenden zu bewältigen sein sollten. Kompetenzniveaumodelle andererseits stellen dar, nach welchen Niveaustufen einzelne Sprachlernende eingeordnet werden können. Gemäß Canale und Swain (1980: 1) kann eine Definition kommunikativer Kompetenz nicht nur zu einem sinnvollen und erfolgreichen Unterricht führen, sondern zudem einen erhöhten Grad von Validität und Reliabilität bei der Bewertung von Sprachfähigkeit ermöglichen. Auch nach Bachman und Palmer, bekannten Testforschern, ist es relevant, die Sprachfertigkeit zu definieren, um Schlussfolgerungen über die Sprachkompetenzen eines Individuums ziehen zu können. Dabei kann auch die Sprachfähigkeit von anderen Faktoren unterschieden werden, die das Testergebnis beeinflussen können (Bachman & Palmer 2010: 43). Das Modellieren von Kompetenzen, und gegebenenfalls von Teilkompetenzen, kann aber auch ein besseres Verständnis für Unterschiede im Hinblick auf quantitative und qualitative Aspekte bei der Evaluierung individueller Leistungen ermöglichen (vgl. Klieme & Leutner 2006), d. h. *welche* und *wie viele* Dimensionen eine Leistung zeigt bzw. auf welchem Niveau jene Leistung die Anforderungen erfüllt haben.

Im Paradigmenwechsel in der Fremdsprachendidaktik von strukturalistischen Ausgangspunkten zu kommunikativen und handlungsorientierten Ansätzen, der in den siebziger Jahren begann, war der Bedarf nach einer Definition der kommunikativen Sprachfähigkeit deutlich zum Vorschein gekommen. Im Bereich des Fremdsprachentestens wurden seitdem mehrere *Kompetenzstrukturmodelle* vorgeschlagen. Von diesen Modellierungen haben insbesondere die Modelle von Canale und Swain (1980; vgl. auch Canale 1983) und von Bachman und Palmer (1996; vgl. auch Bachman 1990) sowie zuletzt das Modell des europäischen Referenzrahmens (vgl. Europarat 2001) eine große Auswirkung gehabt.[49] Im Folgenden werden das weit verbreitete Kompetenzstrukturmodell von Bachman und Palmer (1996) sowie der für die vorliegende Studie zentrale Ansatz des Referenzrahmens (2001) näher beschrieben. Der europäische Referenzrahmen nimmt Bezug auf beide oben genannten Modellierungen, indem er sowohl ein nach Komponenten konzipiertes Modell als auch ein gestuftes Kompetenzmodell enthält. Für das Verständnis von Kompetenz in der vorliegenden Arbeit sind somit beide Arten von Kompetenzmodellierungen relevant.

Zu den bekanntesten Kompetenzstrukturmodellen gehört das Kompetenzmodell von Bachman und Palmer (1996). Dieses Modell ist aber auch auf frühere Vorlagen, wie Bachman (1990), Canale und Swain (1980) und Hymes (1972) zurückzuführen. Unter dem Konzept von *communicative language ability* (CLA) verstehen Bachman und Palmer sowohl linguistische als auch nicht-linguistische Komponenten, die bei der Sprachverwendung miteinander interagieren. Es handelt sich einerseits um die strategische Kompetenz (*strategic competence*), definiert als metakognitive Strategien, die eine effektive und angemessene Sprachverwendung ermöglichen und andererseits um die Sprachkompetenz (*language knowledge*), wie grammatisches und soziolinguistisches Wissen (vgl. Bachman & Palmer 1996). Auch die Sprachkompetenz besteht aus unterschiedlichen Teilen und das von Bachman und Palmer vorgeschlagene Sprachkompetenzmodell enthält folgende Komponenten, siehe Abb. 3:

---

49 Weitere Modelle kommunikativer Kompetenz sind u. a. bei Fulcher und Davidson (2007) aufgeführt. Für einen Vergleich verschiedener Kompetenzmodelle, siehe z. B. McNamara 1996, Harsch 2006 und Lenz 2006. Ein Überblick über die Auseinandersetzung unterschiedlicher Modelle kommunikativer Kompetenz ist u. a. bei McNamara (1996: 48–90) zu finden.

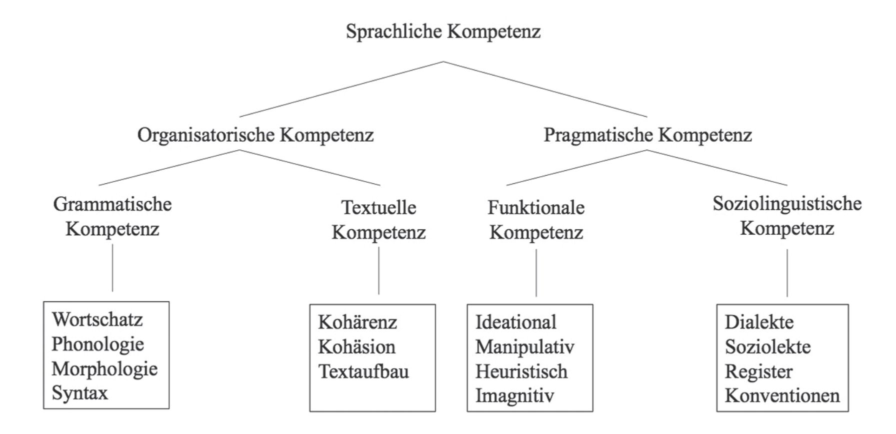


Abb. 3: Komponenten der Sprachkompetenz nach Bachman und Palmer (1996: 68)

Das Modell umfasst eine Unterteilung in die organisatorische Kompetenz und die pragmatische Kompetenz. Darüber hinaus sind folgende Teilkomponenten im Modell enthalten: die grammatische Kompetenz, die Textkompetenz, die funktionale Kompetenz und die soziolinguistische Kompetenz. Bachman und Palmers Modell macht zwar keine Aussagen darüber, wie ein Test gestaltet werden soll, zeigt aber, welche Fähigkeiten entwickelt werden müssen, die später bei der Beurteilung der Sprachkompetenz eines Individuums verwendet und abgeprüft werden können.

Auch wenn die Aufgliederung nach Bachman und Palmers Kompetenzstrukturmodell im Bereich des Fremdsprachentestens weit verbreitet und anerkannt ist, hat das Modell Kritik erhalten, vor allem im Hinblick darauf, dass es sich in der Praxis schwierig umsetzen lässt. McNamara (1996: 75, 85 ff.) weist darauf hin, dass Bachman und Palmers Modell kommunikativer Kompetenz stark vereinfacht sei und dass es interaktionelle Aspekte und die faktische Sprachverwendung nicht zufriedenstellend berücksichtige. Harding (2014: 191) behauptet dahingegen, dass das Modell von Testentwicklern wegen seiner Komplexität eher in verarbeiteten und vereinfachten Formen verwendet werde. Während Kompetenzmodelle wie das von Bachman und Palmer dementsprechend einerseits als nicht komplex genug betrachtet werden, wird andererseits befürchtet, dass sie sich wegen ihrer Komplexität schwierig umsetzen lassen. Nichtsdestoweniger liegt das Modell von Bachman und Palmer vielen der gegenwärtigen Modellierungen und Standards im Bereich des Sprachtestens zugrunde.

Auch der Ansatz im GER basiert auf vorherigen Modellen kommunikativer Kompetenz, wie dem von Bachman und Palmer (1996). Um Kritikpunkte vorheriger Kompetenzmodelle anzugehen, hat der Referenzrahmen einige Erweiterungen im Vergleich z. B. zum Modell von Bachman und Palmer vorgenommen und stellt ein differenziertes Modell zur Beschreibung von Sprachverwendung und Sprachverwendenden dar (vgl. Europarat 2001: Kap. 4). Hierbei wird auch der handlungsorientierte Ansatz des Referenzrahmens deutlich, indem Sprachlernende als angehende Sprachverwendende angesehen werden, aber auch da der Kontext der Sprachverwendung, Themen der Kommunikation, kommunikative Aufgaben und Ziele sowie kommunikative Aktivitäten und Strategien einbezogen werden.

Neben den allgemein definierten Kompetenzen wird die kommunikative Sprachkompetenz beleuchtet (vgl. Europarat 2001: Kap. 5). Zu den Kompetenzbeschreibungen des GER gehören folgende Definitionen:

> Zur Umsetzung ihrer kommunikativen Absichten setzen Sprachverwendende/Lernende sowohl […] ihre allgemeinen Fähigkeiten als auch eine spezifisch sprachbezogene kommunikative Kompetenz ein. Kommunikative Kompetenz in diesem engeren Sinn besteht aus folgenden Komponenten: linguistische Kompetenzen; soziolinguistische Kompetenzen; pragmatische Kompetenzen. (Europarat 2001: 109)

Wie aus dem Zitat ersichtlich, enthält der Referenzrahmen ein Modell fremdsprachlicher Kompetenzen, das mehrere Komponenten der kommunikativen Kompetenz umfasst: *linguistische* Kompetenzen, *soziolinguistische* Kompetenzen und *pragmatische* Kompetenzen. In Abb. 4 sind diese Komponenten und ihre Teilkomponenten aufgeführt:

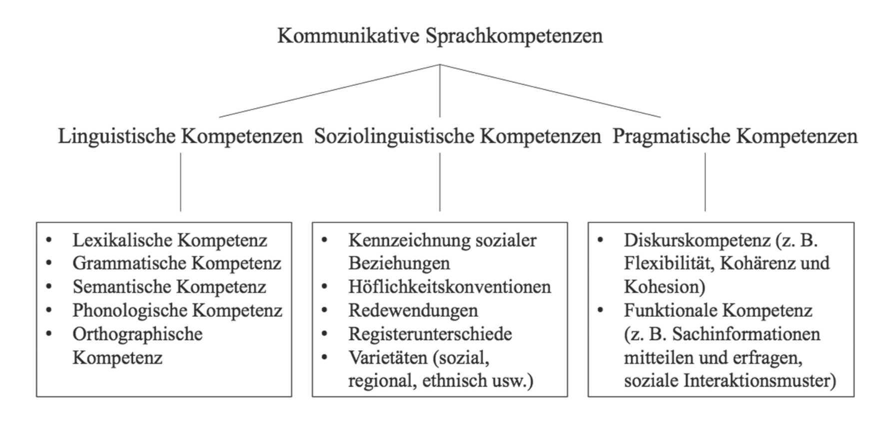


Abb. 4: Komponenten der kommunikativen Sprachkompetenz des GER (Europarat 2001: 109–130)

Auch wenn die Einteilung und die Beschreibungen der Komponenten im Referenzrahmen, vor allem hinsichtlich der linguistischen bzw. soziolinguistischen Kompetenzen, dem Modell von Bachman und Palmer ähneln, gibt es Unterschiede, vor allem im Hinblick auf die pragmatischen Kompetenzen.[50] Die einzelnen Komponenten werden im GER genauer beschrieben, beinhalten jedoch auch unterschiedlich viele Skalen mit gegebenenfalls dazugehörigen Deskriptoren. Die *linguistischen* Kompetenzen sind beispielsweise durch folgende Skalen illustriert: die lexikalische, die grammatische, die phonologische und die orthographische Kompetenz (vgl. Europarat 2001: 110–118). Zu den *soziolinguistischen* Kompetenzen gehören z. B. die Kenntnis und Beachtung von Höflichkeitskonventionen und Registerunterschieden sowie die in der Skala zur soziolinguistischen Angemessenheit erfasste Fähigkeit, sprachliche Variation zu erkennen (vgl. Europarat 2001: 118–122). Im Bereich der *pragmatischen* Kompetenzen existieren unter *Diskurskompetenz* u. a. folgende Skalen: Flexibilität in Bezug auf die Umstände der Kommunikationssituation, Kohärenz und Kohäsion, Flüssigkeit sowie Genauigkeit des Ausdrucks. Auch die Kenntnis verschiedener Gestaltungskonventionen im Hinblick auf die thematische Organisation und die äußere Form geschriebener Texte gehört dieser Kategorie an. Unter *funktionaler Kompetenz* sind zudem Funktionen wie Sachinformationen mitteilen und erfragen sowie soziale Routinen zu finden (vgl. Europarat 2001: 123–130). Des Weiteren teilt der Referenzrahmen in Anlehnung an die traditionelle Gliederung in die vier Fertigkeiten Hörverstehen, Leseverstehen sowie mündlicher und schriftlicher Ausdruck die kommunikativen Sprachaktivitäten in rezeptive, produktive, interaktive und sprachvermittelnde[51]

---

50 In vielerlei Hinsicht sind dieselben Komponenten im Modell der CLA von Bachman und Palmer bzw. der kommunikativen Kompetenz des GER identifiziert worden, diese sind aber teilweise unterschiedlich organisiert. Die grammatikalische Kompetenz in Bachman und Palmers Modell, der organisatorischen Kompetenz zugeordnet, hat ihr Äquivalent in der linguistischen Kompetenz im GER. Während die Textkompetenz in Bachman und Palmers Modell ebenfalls unter organisatorischer Kompetenz eingeordnet wird, ist die ähnlich definierte Diskurskompetenz im Gegensatz dazu unter der pragmatischen Fähigkeit im GER-Modell eingeordnet. Ferner ist im GER die soziolinguistische Kompetenz eine eigene Kategorie, während Bachman und Palmer soziolinguistisches Wissen als eine Untergruppe der pragmatischen Kompetenz betrachten. In beiden Modellen ist die funktionale Kompetenz der Pragmatik zugeordnet. Zusammenfassend betrifft der größte Unterschied somit das, was als pragmatische Fähigkeit definiert wird.

51 Im Begleitband zum GER sind neu herausgearbeitete Skalen zur Sprachvermittlung (*mediation*) zu finden (Council of Europe 2020).

Aktivitäten ein, worunter diese weiter in mündlichen bzw. schriftlichen Sprachgebrauch unterteilt werden können (Europarat 2001: 25–26). Die GER-Skalen zur schriftlichen Produktion bzw. Interaktion allgemein sind in ihrer Ganzheit im Anhang der vorliegenden Arbeit zu finden (vgl. Anhang 3 bzw. 4), was auch für eine Auswahl der hier bereits erwähnten Skalen des GER zutrifft (vgl. Anhang 5–8).

Der Referenzrahmen enthält nicht nur ein System, das unterschiedliche Komponenten der kommunikativen Sprachkompetenz beschreibt, sondern stellt auch ein System mit Kompetenzniveaus, den sechs Referenzniveaus mit den Kürzeln A1–C2, dar. Kompetenzniveaumodelle wie dieses zielen darauf ab, einen Rahmen für die Beschreibung und Beurteilung fremdsprachlicher Kompetenzen, unabhängig von der Sprache, anzubieten. Zusammen mit dem Modell des GER gehören die in den USA herausgearbeiteten Leitlinien sprachlicher Handlungsfähigkeit des American Council on the Teaching of Foreign Languages, ACTFL (2012), im Bereich des Fremdsprachentestens zu den einflussreichsten Kompetenzniveaumodellen.[52] Wie der Referenzrahmen stellen die ACTFL Leitlinien die kommunikative Sprachverwendung in den Mittelpunkt, berücksichtigen Komponenten der Sprachverwendung wie Themen der Kommunikation, Kontexte der Sprachverwendung sowie kommunikative Strategien und enthalten Beschreibungen der kommunikativen Sprachfähigkeit von Lernenden auf unterschiedlichen Kompetenzniveaus. Diese Modellierungen unterscheiden sich aber bezüglich der Anzahl von Kompetenzniveaus. Auch wenn diese beiden einflussreichen Kompetenzniveaumodelle über die Jahre eine weite Verbreitung gefunden haben, ist ihre Verwendbarkeit für Testentwicklung und Testbewertung diskutiert worden (vgl. Fulcher 2016). Trotz Kritik, u. a. im Hinblick auf mangelnde empirische Forschung hinsichtlich der unterschiedlichen Referenzniveaus (vgl. Alderson 2007; Chalhoub-Deville 1997; Hulstijn 2007; Wisniewski 2014), werden Kompetenzniveaumodellierungen, darunter vor allem der Referenzrahmen, zunehmend als eine Basis für Leistungsmessungen von Sprachlernenden verwendet.

---

52 Die ACTFL Leitlinien sprachlicher Handlungsfähigkeit, erstmal in den 80er Jahren veröffentlicht und seitdem einige Male überarbeitet, stellen für die vier klassischen Fertigkeiten Hören, Lesen, Schreiben und Sprechen fünf Leistungsniveaus bereit: *Distinguished*, *Superior*, *Advanced*, *Intermediate* und *Novice*. Die letzteren drei sind weiterhin in drei Subniveaus unterteilt, *High*, *Mid* und *Low*. In den letzten Jahren hat eine zunehmende Anzahl von Studien das Verhältnis zwischen Ergebnissen, die zu den jeweiligen Kompetenzniveaus der ACTFL Leitlinien und des GER gehören, untersucht (siehe z. B. Tschirner & Bärenfänger 2012 für Deutsch als Fremdsprache).

Das gestufte Kompetenzniveaumodell des Referenzrahmens liegt den Bildungsstandards für Fremdsprachen in mehreren Ländern zugrunde. Das vorliegende System für die Fremdsprachen in Schweden ist als *Basisstandards* konstruiert, was beinhaltet, dass das Mindestniveau der jeweiligen Fremdsprachenstufen an den entsprechenden Kompetenzniveaus des GER orientiert ist, z. B. die ausreichende Note E auf der Fremdsprachenstufe *Tyska 5* an dem B1-Niveau (vgl. Skolverket 2011b). Die jeweiligen Niveaustufen des GER bieten somit, wie in der vorliegenden Studie, einen Referenzpunkt für empirische Untersuchungen im Hinblick auf das Erfüllen der Mindestanforderungen an Sprachkompetenzen bei Schülerinnen und Schülern. Die Bildungsstandards in Ländern wie z. B. Deutschland sind hingegen *Regelstandards*, was bedeutet, dass diese sich stattdessen nach einem durchschnittlichen Anforderungsniveau orientieren (Klieme et al. 2003).

Da das Mindestniveau der Anforderungen für die Fremdsprachenstufe *Tyska 5* sich gemäß der schwedischen Schulbehörde an dem GER-Niveau B1 orientiert, sind die Skalen dieser Niveaustufe für die vorliegende Arbeit besonders relevant. Deskriptoren, die kommunikative Fertigkeiten auf einem Niveau B1 beschreiben, sind in mehreren verschiedenen Skalen des GER zu finden. Zur Illustration sind die Skalen zur Erfassung der schriftlichen Produktion bzw. Interaktion allgemein in Tab. 7 und Tab. 8 aufgeführt:

Tab. 7: *GER-Skala des B1-Niveaus für schriftliche Produktion allgemein (*Europarat 2001*: 67)*

GER-Niveau B1	*Schriftliche Produktion allgemein* Kann unkomplizierte, zusammenhängende Texte zu mehreren vertrauten Themen aus seinem/ihrem Interessengebiet verfassen, wobei einzelne kürzere Teile in linearer Abfolge verbunden werden.

Tab. 8: *GER-Skala des B1-Niveaus für schriftliche Interaktion allgemein (*Europarat 2001*: 86)*

GER-Niveau B1	*Schriftliche Interaktion allgemein* Kann Informationen und Gedanken zu abstrakten wie konkreten Themen mitteilen, Informationen prüfen und einigermaßen präzise ein Problem erklären oder Fragen dazu stellen. Kann in persönlichen Briefen und Mitteilungen einfache Informationen von unmittelbarer Bedeutung geben oder erfragen und dabei deutlich machen, was er/sie für wichtig hält.

Insgesamt beschreiben diese Skalen das Ziel eines Lernprozesses und stellen zugleich Anforderungen im Hinblick auf das Ergebnis jenes Lernprozesses, d. h. was ein Sprachlernender bezüglich der schriftlichen Produktion und Interaktion auf jenem Niveau bewältigen muss. In den Skalen werden sprachliche Handlungen durch konkrete Kann-Beschreibungen vorgelegt, die das Niveau für sowohl Sprachlernende als auch Sprachlehrende transparent und anschaulich zu erklären versuchen. Die Beschreibungen versuchen dabei ein breites Bild der schriftlichen Sprachkompetenz eines Sprachlernenden darzustellen. Um Sprachhandlungen auf einem bestimmten Niveau bewältigen zu können, müssen die Sprachlernenden nach dem Modell des GER Zugang zu verschiedenen Kompetenzen haben. Dies bedeutet nicht nur die Bewältigung einer Reihe von Sprachkompetenzen, wie Wortschatzbeherrschung, soziolinguistische Angemessenheit sowie Kohärenz und Kohäsion, die in den drei Hauptkomponenten der kommunikativen Sprachkompetenz des GER einzuordnen sind (vgl. Abb. 4), sondern auch, dass Lernenden allgemeine Kompetenzen wie Weltwissen und soziokulturelles Wissen zur Verfügung stehen.

Die Nachfrage nach einer Orientierung von Sprachtests und Lernergebnissen an externen Kompetenzniveaumodellen, häufig als Standards und Rahmenwerke[53] bezeichnet, kann mit einem erhöhten Fokus auf Verantwortlichkeit (*accountability*) im Bildungsbereich in Verbindung gesetzt werden (z. B. Chapelle et al. 2020). Auch im Bereich Fremdsprachentesten hat in letzter Zeit das Interesse für die Anbindung von Bildungsstandards, Lehrbüchern und

53 Die Begriffe *Standards* oder *Rahmenwerke* werden heute häufig synonym verwendet. Ursprünglich stammt die Verwendung des Begriffs *Standards* aus dem kriterienbasierten Testparadigma, wonach Leistungen gemäß vordefinierten Kriterien anstatt nach der relativen Position einer Skala interpretiert wurden. Inzwischen wird der Begriff *Standards* oft in Zusammenhang mit Dokumenten wie den ACTFL Leitlinien sprachlicher Handlungsfähigkeit oder dem GER verwendet, nicht zuletzt, wenn Testergebnisse zu externen Standarddokumenten zugeordnet werden sollten (vgl. Fulcher 2016). Ein Rahmenwerk ist nach Fulcher und Davidsson (2007: 36): „a selection of skills and abilities from a model that are relevant to a specific assessment context“ und vermittelt somit zwischen einem Modell und Testspezifikationen, z. B. hinsichtlich Inhalte und Formate der Aufgaben. Inwiefern der GER als ein Standard, ein Rahmenwerk oder ein Modell zu verstehen ist, kann allerdings diskutiert werden. Ein Modell sprachlicher Kompetenz legt gemäß Fulcher und Davidson (2009: 126) eine theoretische Beschreibung darüber dar, was es bedeutet, eine Sprache zu beherrschen und zu verwenden – eine Definition, die nach Fulcher und Davidson auf den GER zutrifft. Da sämtliche drei Begriffe häufig im Zusammenhang mit dem GER vorkommen, werden sie in der vorliegenden Arbeit gleichermaßen verwendet.

insbesondere standardisierten Sprachprüfungen an externe Modellierungen stark zugenommen. Häufig definieren und beschreiben diese Rahmenwerke unterschiedliche Leistungsniveaus oder Inhalte, die zu einem bestimmten Niveau gehören. Als Beispiele solcher Rahmenwerke sind die bereits oben erwähnten ACTFL Leitlinien sprachlicher Handlungsfähigkeit (American Council on the Teaching of Foreign Languages 2012), Canadian Language Benchmarks, CLB (Centre for Canadian Language Benchmarks 2019), und der GER (Europarat 2001) zu nennen. Innerhalb Europas wird vor allem der GER als Bezugspunkt von Anbindungsstudien verwendet (vgl. McNamara & Roever 2006).

Dass die Anbindung von Sprachtests bezüglich der Orientierung an externen Kompetenzstandards bisher im Mittelpunkt stand, zeigt nicht zuletzt die Veröffentlichung eines Manuals, das den Anbindungsprozess von Sprachtests an die jeweiligen Referenzniveaus des GER beschreibt. Der weit verbreitete Trend, bereits vorhandene *High-Stakes-Tests* dem GER zuzuordnen, hat seitdem dazu geführt, dass heute viele der wichtigsten Sprachtests in Europa und zunehmend weltweit auf den GER ausgerichtet sind (Harsch & Hartig 2015: 334).

Viele der bisherigen Studien zur Validierung standardisierter Sprachtests haben das vom Europarat vorgeschlagene methodische Verfahren nach dem publizierten Manual (vgl. Kap. 2.3.2) geprüft und evaluiert. Auch wenn Bildungsstandards in den Mitgliedsländern der Europäischen Union sich nach Beschluss des Europaparlaments an den Deskriptoren und Referenzniveaus des GER orientieren sollten (Council of Europe 2008), haben weniger empirische Studien die Sprachkompetenzen von Schülerinnen und Schülern in einer Fremdsprache ausgewertet und diese gleichzeitig auf den Europäischen Referenzrahmen sowie auf die jeweiligen Bildungsstandards in den Ländern bezogen (vgl. Kap. 4.2). Ein ähnliches Handbuch wie das *Manual* mit Richtlinien zur Validierung von Sprachleistungen in Bezug auf den GER, das ein methodisches Verfahren zur Sprachstandfeststellung unter Lernenden darstellt, existiert jedoch noch nicht. Für Sprachleistungsstudien mit Bezug auf den GER sind daher bislang zum Teil unterschiedliche methodische Ansätze zum Einsatz gekommen.

3.2 Validität

In diesem Kapitel wird zu Beginn im ersten Teil kurz auf die historische Entwicklung und Eingrenzung des Validitätskonzepts eingegangen. Zunächst folgt darauf eine Beschreibung der heute im Bereich des Fremdsprachentestens dominierenden Definitionen zum Begriff Validität von Messick (1989b). Im

Hinblick auf die Relevanz für die vorliegende Studie werden danach im zweiten Teil zwei Validierungsmodelle, der argumentbasierte Ansatz (vgl. Bachman 2005; Bachman & Palmer 2010; Kane 2006; 2013; Chapelle et al. 2008; Chapelle 2020) bzw. das soziokognitive Rahmenmodell (vgl. Weir 2005; Shaw & Weir 2007; O'Sullivan & Weir 2011), dargestellt.

3.2.1 Entwicklung und Begriffseingrenzung des Validitätskonzepts

Das Konzept der Validität ist im Laufe der letzten 80 Jahre unterschiedlich verstanden und definiert worden. Eine einfache Definition von Validität lautet, dass ein valider Test das misst, was er zu messen versucht (z. B. Cureton 1951). In den 1940er Jahren wurden verschiedene Methoden, um Validität festzustellen, vorgeschlagen, aus denen heraus sich über die Jahrzehnte eine Reihe von unterschiedlichen Validitätstypen und Definitionen entwickelten (vgl. Newton & Shaw 2014). Nach moderneren Ansätzen geht man von einem einheitlichen Validitätskonzept aus, wonach Validität sich auf zugehörige Interpretationen von Testergebnissen bezieht, deren Plausibilität vom Anwendungskontext abhängig ist. Heute spricht man dementsprechend eher davon, dass eine Sprachprüfung valide ist, wenn man angemessene und nützliche Schlussfolgerungen ziehen kann. Validität wird somit nicht, wie oft die bisherige Auffassung, als ein Gütekriterium des Prüfungsverfahrens oder faktischen Tests verstanden, sondern bezieht sich eher auf die Bedeutung der Testergebnisse (vgl. Messick 1989a; Kane 2006; Bachman & Palmer 2010). Von einer einheitlichen Entwicklung kann nicht gesprochen werden, auch wenn heute die Perspektive eines einheitlichen Validitätskonzepts dominierend ist. Im heutigen Forschungsfeld existieren dementsprechend eine Vielfalt von Auffassungen und Definitionen, die mehr oder weniger voneinander abweichen.[54]

Ein gewisses Maß an Übereinstimmung über gegenwärtige Validitätskonzepte, Methoden und Wertimplikationen haben die in den USA für Testzwecke erschienenen *Standards for educational and psychological tests and manuals*[55]

---

54 Vgl. hierzu z. B. Borsboom et al. (2004), die eine einfachere Definition von Validität im Vergleich zu gegenwärtigen Modellen vorschlagen und Validität als eine Eigenschaft des Tests verstehen. Das Validitätskonzept wird eher traditionell definiert: „It [validity] is a very basic concept and was correctly formulated, for instance, by Kelley (1927, p. 14) when he stated that a test is valid if it measures what it purports to measure" (Borsboom et al. 2004: 1061).

55 Diese Dokumente werden in der Regel als die *Standards* abgekürzt. In der ersten Version aus dem Jahr 1954 mit dem Titel *Technical recommendations for psychological tests and diagnostic techniques,* von der *American Psychological Association*, APA, herausgegeben, aber seit 1966 und bei den späteren Ausgaben der *Standards* aus den

geschaffen. Eine wichtige Rolle bei der Entwicklung und Begriffseingrenzung des Validitätsbegriffs spielt zudem das Validitätskapitel in dem einschlägigen und in mehreren Auflagen erschienenen Handbuch *Educational Measurement* (vgl. Messick 1989b; Kane 2006). Die *Standards*, oft beeinflusst durch das vorher erschienene Kapitel in *Educational Measurement,* können somit gleichzeitig auch als Anzeichen für Veränderungen hinsichtlich des Validitätskonzepts angesehen werden.[56]

Traditionell sind drei Typen von Validität definiert worden: Inhaltsvalidität (*content validity*), kriterienbezogene Validität (*criterion validity*) und Konstruktvalidität (*construct validity*), jeweils mit bestimmten Aspekten von Validität verbunden (vgl. Messick 1989a). *Inhaltsvalidität* bezieht sich darauf, inwiefern der Testinhalt eine repräsentative Auswahl des zu testenden Domänenbereichs widerspiegelt. Die *kriterienbezogene Validität* umfasst Vergleiche zweier Variablen, wobei die eine Variable das Testergebnis ist. Die kriterienbezogene Validität gibt an, inwieweit das Testergebnis mit einem oder mehreren externen empirischen Aspekten (sog. Kriterien), von denen angenommen werden kann, dass sie vom Test geprüft werden sollen, oder mit anderen Testergebnissen in Verbindung steht. Wenn wir dabei aus den Testergebnissen korrekte Schlüsse über zukünftige Leistungsfähigkeit in Form von zukünftigen Kompetenzniveaus oder akademischem Erfolg ziehen können, wird von Vorhersagevalidität (*predictive validity*) gesprochen. Wenn wir dahingegen gleichzeitig vorliegende Ergebnisse, wie z. B. Testergebnisse von einer mündlichen bzw. schriftlichen Prüfung, miteinander vergleichen, wird von Übereinstimmungsvalidität (*concurrent validity*) die Rede sein. Abschließend wird zur Bestimmung der *Konstruktvalidität*[57] evaluiert, inwieweit ein Test die zu messenden Merkmale oder Eigenschaften, d. h. das Konstrukt[58], tatsächlich abprüft. Konstrukte, auch als

---

Jahren 1974, 1985, 1999 und 2014 auch von der *American Educational Research Association*, AERA, und dem *National Council on Measurement in Education*, NCME.

56 Vgl. Chapelle (2020: 14–15) für einen Überblick darüber, wie sich das Validitätskonzept z. B. im Hinblick auf Reliabilität, Konstrukt, Testverwendung, Wertimplikationen und Konsequenzen in den verschiedenen Auflagen von *Educational Measurement* verändert und entwickelt haben.

57 Cronbach und Meehls Definition der Konstruktvalidität aus dem Jahr 1955 hat eine zentrale Rolle für die Konturierung des Begriffes gespielt und sollte dann verwendet werden, wenn das zu messende Attribut oder die zu messende Qualität nicht operationalisiert ist oder sich nicht als ein präzises Kriterium definieren lässt (vgl. Cronbach & Meehl 1955).

58 Ein Testkonstrukt kann unterschiedlich definiert werden. Chapelle (1998) unterscheidet hierbei, u. a. basierend auf Messick (1989b), zwischen verschiedenen

„traits“ oder „Attribute“ bezeichnet (vgl. Kane 2006: 30 ff.; Newton & Shaw 2014: 11 ff.), werden von Weir (2005) als „underlying […] abilities we wish to measure in students, the *what* of language testing“ (S. 1) definiert. So sollen beispielsweise Vergleiche der Messergebnisse von Testverfahren, die dasselbe Konstrukt realisieren, hoch miteinander korrelieren, als sog. *konvergente Validität* bezeichnet, während das umgekehrte Verhältnis bei der sog. *diskriminanten Validität* vorliegt, was bedeutet, dass Messergebnisse, die unterschiedliche Konstrukte wiedergeben, eine geringere Korrelation aufzeigen.

Ein validitätstheoretischer Beitrag von Messick (vgl. 1989b) hat für das heutige Verständnis von Validität einen großen Einfluss gehabt. Messicks eigene Definition der Validität lautet:

> Validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *inferences* and *actions* based on test scores or other modes of assessment (1989a: 5, *Hervorheb. im Original*)

Anhand dieser Definition soll durch ein integriertes Urteil ermittelt werden, *zu welchem Grad* wir Inferenzen aus einem Testergebnis ziehen können (z. B. über die sprachliche Kompetenz eines Lernenden) und inwiefern die dazugehörigen Beschlüsse berechtigt sind (z. B. Zulassung zu einem bestimmten Studiengang). Diese Konzeptualisierung von Validität als ein einheitliches Konzept ist durch Messick bekannt geworden und wurde weitgehend angenommen. Gemäß Messick sind die bislang als gültig angesehenen trennbaren Validitätstypen, die unterschiedliche Arten von Nachweisen benötigen, jeder für sich nur unter bestimmten Umständen relevant: „neither content nor criterion validity alone is sufficient for any testing purpose“ (1989a: 6). Er beschreibt an dieser Stelle die Konstruktvalidität als das zentrale Konzept für die Validität:

Definitionsansätzen eines Testkonstruktes: traitzentrierte Ansätze, behavioristische Ansätze und interaktionale Ansätze. In einem traitzentrierten Ansatz (*trait approaches*) bezieht sich das Konstrukt auf die Fähigkeiten und Eigenschaften der Lernenden, die zur Bewältigung der Testaufgabe gebraucht werden. Nach einem behavioristischen Ansatz (*behaviorist approaches*) liegt wiederum der Fokus auf kontextuellen Faktoren, z. B. inwieweit die Lernenden Aufgaben hinsichtlich künftiger Verwendungskontexte bewältigen können. Gemäß der Definition interaktionaler Ansätze (*interactionalist approaches*) bilden „traits, contextual features, and their interaction“ (vgl. Chapelle 1998: 34), d. h. die Interaktion hinsichtlich Fähigkeiten und Eigenschaften der Lernenden, kontextuellen Faktoren und der Wechselwirkung zwischen ihnen, den Rahmen für die Konstruktdefinition.

> Validity is broadly defined as nothing less than an evaluative summary of both the evidence for and the actual – as well as potential – consequences of score interpretation and use (i.e., construct validity conceived comprehensively). This comprehensive view of validity integrates considerations of content, criteria, and consequences into a construct framework for empirically testing rational hypotheses about score meaning and utility. (Messick 1995: 742).

Messick betont hierbei, dass sich Validität auf die Interpretation und Verwendung der Testergebnisse und nicht auf den Test selbst bezieht. Konstruktvalidität handelt demzufolge nicht von der jeweiligen Prüfung, sondern davon, in welchem Umfang wir die Testergebnisse als Attribute für das, was wir messen wollen, interpretieren können. Diese Definition der Validität wird u. a. auch später von Kane unterstützt: „The distinction here is not among different kinds of validity or even different types of validity evidence, but among different types of interpretations“ (Kane 2001: 334). Validität ist demnach facettenreich und verlangt verschiedene Typen von Nachweisen. Diese moderne Definition von Validität legt somit deutlich den Fokus auf die Interpretation und die Verwendung der Testergebnisse, deren Argumente sowohl eine qualitative als auch eine quantitative Basis haben können (vgl. Chapelle 2020).

Die Validität eines Tests erfordert gemäß Messick (1989a) eine umfassende Auswertung, d. h. eine Validierung, die verschiedene Facetten des Validitätskonzepts berücksichtigt. Messick definiert Validierung[59] als einen Prozess, in welchem Argumente über Interpretationen und die Verwendung von Testergebnissen gesammelt werden, die durch theoretische Begründungen und empirische Beweise unterstützt werden können. In einem Modell von Messick (ibid.) werden verschiedene Facetten der Validität abgebildet und hierbei werden zwei miteinander verbundene Dimensionen des einheitlichen Vadilitätskonzepts aufgestellt:

---

59 Das Validitätskapitel im *Educational Measurement* wird abwechselnd *validity* (Validität) bzw. *validation* (Validierung) benannt (vgl. Cureton 1951; Cronbach 1971; Messick 1989b; Kane 2006). Die Bezeichnungen Validität und Validierung sind miteinander eng verbunden. Die Validität bezieht sich aber auf den Begriff, während die Validierung eine Qualitätssicherung ist und sich auf den Prozess, worin logische Argumente und empirische Nachweise hinsichtlich des Validitätskonzepts evaluiert werden, bezieht.

Tab. 9: *Struktur der Validitätsfacetten (nach* Messick 1989a*: 10)*

Quelle der Rechtfertigung	*Testinterpretation*	*Testverwendung*
Evidentielle Basis	1. Konstruktvalidität	2. Konstruktvalidität + Relevanz/ Nützlichkeit
Konsequenzielle Basis	3. Konstruktvalidität + Wertimplikationen	4. Konstruktvalidität + Relevanz/ Nützlichkeit + Wertimplikationen + soziale Konsequenzen

Das Modell differenziert zwischen einerseits der Funktion eines Tests, d. h. der *Testinterpretation* und der *Testverwendung*, und andererseits der *Quelle der Rechtfertigung* für die Bewertung, basierend auf Nachweisen, die bedeutungstragend für die Testergebnisse sind (*evidentielle Basis*) oder basierend auf Werten und Konsequenzen, die zur Werteinschätzung des Tests beitragen (*konsequenzielle Basis*). Messick hebt jedoch hervor, dass die verschiedenen Komponenten in den Zellen oben nicht nur miteinander verbunden sind, sondern auch überlappen, was nach Messick daran liegt, dass hier versucht wurde, einzelne Teile aus einem zusammenhängenden Konzept darzustellen. Die evidentielle Basis zur Evaluation der Testinterpretation in der ersten Zelle ist Konstruktvalidität in Form von empirischen Nachweisen und theoretischen Analysen. Auch zur Evaluation der Testverwendung, in der zweiten Zelle oben, ist die evidentielle Basis die Konstruktvalidität, aber hinzu kommt die Relevanz und Nützlichkeit für den spezifischen Kontext. Hierbei können wir uns fragen, inwiefern der Test für eine bestimmte Lerngruppe in einem spezifischen Lernkontext geeignet ist. Die konsequenzielle Basis für Testinterpretationen, in der dritten Zelle, bezieht sich neben der Konstruktvalidität auch auf Wertimplikationen. In diesem Fall wird untersucht, inwiefern die Testinterpretation angesichts der damit verbundenen Werte angemessen ist. Abschließend liegt die konsequenzielle Basis für die Testverwendung in der vierten Zelle neben der Konstruktvalidität, der Relevanz und Nützlichkeit sowie den Wertimplikationen auch in den sozialen Konsequenzen. Hierbei soll u. a. evaluiert werden, welche Entscheidungen durch die Testverwendung getroffen werden und was diese veranlasst hat.

Wenn im Laufe der Zeit neue theoretische und/oder empirische Belege auf evidenzieller Basis entwickelt oder dargestellt werden, würde dies folglich bedeuten, dass die Testinterpretation und/oder die Testverwendung neu bewertet werden muss und diese neuen Nachweise könnten zu einer geringeren

Validität führen (vgl. hierzu z. B. Kane 2013). Darüber hinaus können auch die Interpretation und die Verwendung von Testergebnissen zu geringerer Validität führen, wenn auf konsequenzieller Basis negative Wertimplikationen oder Konsequenzen der Testergebnisse einbezogen werden. Dies könnte der Fall sein, wenn ein Test nicht die beabsichtigten Konsequenzen gibt, z. B. bei akademischem Studienerfolg, auch wenn sonstige Anforderungen erfüllt sind. Inwiefern Dimensionen wie soziale Konsequenzen und Werte überhaupt zum Validitätskonzept gehören sollten oder nicht, ist allerdings in der Forschung umstritten (vgl. Popham 1997; Mehrens 1997; Borsboom et al. 2004; McNamara 2006; McNamara & Roever 2006). Wenn wir ein Testergebnis mit einer Bedeutung füllen, tragen nach Messick die in der Testinterpretation impliziten Werte und die Auswirkungen auf die Gesellschaft sowie die impliziten Werte der Testverwendung zur Konstruktvalidität bei (1995: 748). Andere Forscher sind dagegen der Meinung, dass Konsequenzen zwar als ein Teil der Testqualität angesehen werden sollten, jedoch nicht dem Validitätskonzept angehören (vgl. Shadish et al. 2002: 475 ff.; Kunnan 2004; Lissitz & Samuelson 2007).

Es bestehen nach Messick (vgl. 1989a; 1995) zwei bedeutende Risiken (*threats*) für die Konstruktvalidität eines Tests, die in der Prüfungssituation vermieden werden sollten: die Unterrepräsentation des Konstrukts (*construct underrepresentation*) und die konstruktirrelevante Varianz (*construct-irrelevant variance*). Im Fall einer Unterrepräsentation des Konstrukts ist der Test zu eng gefasst und enthält nicht alle Dimensionen des zu messenden Konstrukts. Dies ist beispielsweise der Fall bei der großen Sprachleistungsstudie ESLC (vgl. European Commission 2012b), die darauf abzielte, die sprachliche Kompetenz unter Jugendlichen auszuwerten, und dabei die mündliche Kompetenz der Lernenden nicht geprüft hat. Ein weiteres Beispiel wäre die Überbetonung grammatischer Korrektheit bei der Bewertung von Textproduktionen. Wenn man die schriftliche Kompetenz als aus mehreren verschiedenen Teilen bestehend versteht, ist die grammatische Korrektheit allein nicht als ausreichender Indikator für die schriftliche Fähigkeit eines Lernenden zu betrachten. Bei einer konstruktirrelevanten Varianz umfasst der Test dahingegen auch Dimensionen, die gemäß dem zu messenden Konstrukt nicht angebracht sind, was dabei gewisse Gruppen von Lernenden systematisch benachteiligt. Beispiele für konstruktirrelevante Varianz sind Aufgaben zum Hörverstehen, die zusätzlich ein umfangreiches Weltwissen von den Lernenden verlangen, Bewertungskriterien, die auch andere, irrelevante Dimensionen miteinbeziehen oder Bewertende, die bei der Beurteilung schriftlicher Kompetenz Textproduktionen mit einer schönen Handschrift höher einstufen.

3.2.2 Validitätsmodelle

Die Struktur des einheitlichen Validitätskonzepts im Rahmenmodell von Messick gilt aber als komplex und bietet nach Kane keine konkreten Richtlinien zur Validierung von Interpretation und Verwendung der Testergebnisse (Kane 1992). Da sie sich in der Praxis schwer umsetzen lässt, sind Modelle vorgeschlagen worden, die nicht nur in der Forschung, sondern auch für die konkrete Verwendung von Praktikern genutzt werden könnten, z. B. das Modell der Testnützlichkeit (*test usefulness*) von Bachman und Palmer (1996), das Modell von Testfairness nach Kunnan (2004) und argumentbasierte Rahmenmodelle (vgl. Kane et al. 1999; Kane 1992; Kane 2001). In den letzten Jahrzehnten haben sich daraus Rahmenmodelle zur Validierung entwickelt, die in systematischer Art und Weise die Gültigkeit verschiedener Thesen nachweisen. In diesen unterschiedlichen Validierungsmodellen wird u. a. definiert, welche Typen von Nachweisen für unterschiedliche Inferenzen zur Testinterpretation und Testverwendung verwendet werden müssen. Diese Methoden und Modelle zur Validierung, wie der argumentbasierte Ansatz von Kane (2006; 2013, vgl. hierzu auch Crooks et al. 1996 bzw. Kane et al. 1999) und das soziokognitive Rahmenmodell (*socio-cognitive framework*) von Weir (2005), werden vor allem im Bereich Testevaluation und Testentwicklung verwendet. Validierungsmodelle werden nicht zuletzt im Hinblick auf die Überprüfung von sog. *High-Stakes*-Prüfungen genutzt, um Qualitätsanforderungen zu begegnen sowie die Interpretation und die Verwendung von Testergebnissen legitimieren zu können. Im Folgenden werden die beiden oben erwähnten argumentbasierten Rahmenmodelle dargestellt, die für die spätere Analyse und Diskussion in der vorliegenden Arbeit von Relevanz sind.

3.2.2.1 Argumentbasierte Ansätze nach Kane

Argumentbasierte Ansätze zur Testvalidierung entwickeln die einheitliche Definition der Validität von Messick weiter (vgl. Kane 1992; Kane et al. 1999). Parallel hierzu haben sich daraus andere argumentbasierte Modelle entwickelt, wie z. B. das nachweisbasierte Testdesign (*Evidence-centered Assessment design*) von Mislevy (Mislevy et al. 2002; Mislevy & Riconscente 2006) und das *Assessment Use Argument* (AUA) von Bachman (2005; Bachman & Palmer 2010). Argumentbasierte Modellierungen (z. B. Kane 2002; 2006; 2013; Chapelle et al. 2008; 2010) sind seitdem im Bereich des Fremdsprachentestens dominierend und sehr einflussreich geworden.

Argumentbasierte Ansätze bieten eine Struktur, um Nachweise der Validität untersuchen zu können. Basierend auf den Theorien zur Argumentationsstruktur

von Toulmin (1958) ist bei Kane eine Argumentationskette, die die Verbindung zwischen der beobachteten Testleistung und der Interpretation der Testergebnisse zeigt, dargestellt. Die Bildung eines Validitätsarguments, bestehend aus logischen Analysen und empirischen Nachweisen, soll dazu dienen, die Adäquatheit und Plausibilität vorgeschlagener Intentionen und Verwendungen zu evaluieren, u. a. im Hinblick auf theoretische Konstrukte, Schlussfolgerungen einer Bewertung und die Konsequenzen. Kane gibt dabei an: „To validate an interpretation or use of measurements is to evaluate the rationale, or argument, for the claims being made" (2006: 17).

Kane identifiziert zwei zentrale Argumente für den Validierungsprozess und konzeptualisiert Validierung als bestehend aus einem interpretativen Argument (*interpretive argument*)[60] und einem Validitätsargument (*validity argument*). Gemäß Kane geschieht die Validierung somit in einem zweistufigen Prozess (vgl. Kane 2006; 2013). In einem ersten Schritt wird das interpretative Argument spezifiziert, was dazu dient, eine Struktur der geplanten Interpretationen und Verwendungen darzulegen:

> An *interpretive argument* specifies the proposed interpretations and uses of assessment results by laying out a network of inferences and assumptions leading from the observed performances to the conclusions and decisions based on the assessment scores (Kane 2011: 8, *Hervorheb. im Original*).

In einem zweiten Schritt wird ein Validitätsargument konstruiert, wodurch logische Analysen und empirische Nachweise eingeholt werden, um die Stärke des interpretativen Arguments evaluieren zu können. Hierbei werden Schlussfolgerungen hinsichtlich ihrer Kohärenz und Plausibilität ausgewertet, wobei Entscheidungen im Hinblick auf ihre Auswirkung oder auf ihre Konsequenzen evaluiert werden (Kane 2006: 51). Kanes Argumentationskette illustriert die Verlinkung von der Beobachtung einer Leistung (*observation*) zu den Entscheidungen (*decisions*). Die einzelnen Glieder der Argumentationskette sind in Abb. 5 aufgeführt:

60 Kane hat diesen Begriff später durch das sog. *Interpretation/Use Argument* (IUA) ersetzt, um nicht nur der Interpretation, sondern auch der Verwendung Gewicht zu geben (Kane 2013: 2).

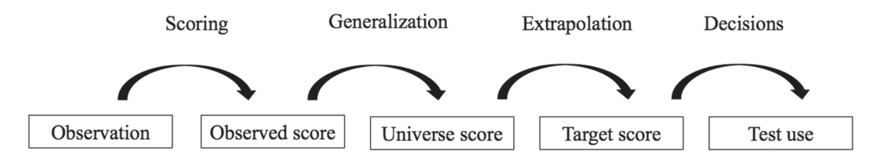


Abb. 5: Darstellung einer Argumentationskette nach Kane (2013)

Kane beschreibt folglich Schritte, nach denen Schlussfolgerungen (*inferences*) gezogen werden können. Die erste Schlussfolgerung, nämlich die Bewertung (*scoring*), bezieht sich darauf, wie die Leistung eines Testteilnehmenden in einem beobachteten Testergebnis (*observed score*) realisiert wird, z. B. bei einer Punktzahl oder einer Note. Hierbei wird u. a. davon ausgegangen, dass bei der Bewertung der Leistung adäquate Kriterien zur Verfügung stehen und dass diese Kriterien wie vorgesehen verwendet werden.

Die zweite Schlussfolgerung, die Generalisierung (*generalization*), befasst sich mit der Frage, inwiefern die Testergebnisse eines bestimmten Tests als Einschätzung (*universe score*) auch auf andere vergleichbare Testsituationen und Testaufgaben zu generalisieren sind. Nach Kane (2006) sollte hierbei generell erwartet werden können, dass die Aufgaben des Tests eine repräsentative Auswahl im Hinblick auf den Zielbereich (*universe of generalization*) bilden – „typically drawn from a subset of the target domain“ (S. 31) – und dass die Testteilnehmenden unter ähnlichen Bedingungen eine ähnliche Aufgabe mit ähnlichem Erfolg bearbeiten können. Die Testergebnisse eines Testteilnehmenden würden jedoch in diesen ähnlichen Parallelformen des Tests eine gewisse Variabilität zeigen. Die Präzision der Generalisierbarkeit der Testergebnisse vom beobachteten Testergebnis zu einem übertragenen Testergebnis, *universe score*, ist von dieser Variabilität begrenzt. Die Variabilität kann aber in verschiedenen Untersuchungen evaluiert werden, z. B. durch Untersuchungen der Bewerterübereinstimmung oder durch sog. IRT-Analysen, durch die eine Einschätzung der Schwierigkeitsgrade der Aufgaben oder Milde-Strenge-Tendenzen der Bewertenden gezeigt werden können.

Die dritte Schlussfolgerung, Extrapolation (*extrapolation*), bezieht sich darauf, dass man das Testergebnis als einen Indikator für die Kompetenz oder die Leistung eines Testteilnehmenden (*target score*) in einer Realsituation sehen kann. Bei der Extrapolation können Annahmen aus den Testergebnissen darauf übertragen werden, inwiefern die Testteilnehmenden ähnliche Aufgaben auch in einer Realsituation leisten können und inwieweit sie damit das Sprachniveau hinsichtlich der Zieldomäne erreicht haben oder nicht. Diese Schlussfolgerung kann u. a. durch logische Analysen oder Korrelationen mit einem

auf dem gleichen Konstrukt basierenden externen repräsentativen Kriterium verglichen werden (Kane et al. 1999).

Im letzten Schritt, Entscheidungen (*decisions*), wird evaluiert, inwiefern die Testergebnisse genügend Informationen über die Kompetenzen der Testteilnehmenden für eine sinnvolle Nutzung (*test use*) geliefert haben, z. B. als Grundlage einer Entscheidung über das sprachliche Niveau, das für einen bestimmten Studiengang gebraucht wird. Alle Schlussfolgerungen stellen das interpretative Argument dar, wobei das beobachtete Testergebnis mit Kompetenzen in der Zieldomäne in Verbindung gesetzt wird. Welche und wie viele logische und/oder empirische Belege werden gebraucht, um die Validität nachweisen zu können? Im Unterschied zu Messick, für den im Großen und Ganzen fast jede Art vom Nachweis oder Analyse zur Validität von Bedeutung ist (vgl. Messick 1989b), sollten nach Kane eigentlich nur die für das interpretative Argument relevanten Behauptungen evaluiert werden:

> The kinds of validity evidence that are most relevant are those that support the main inferences and assumptions in the interpretive argument, particularly those that are most problematic. Conclusions about validity are always tentative in the sense that new evidence or new insights could force a change, but one can get to the point that a proposed interpretation or use is clearly justified, because its inferences and assumptions are supported by empirical evidence and/or are highly plausible *a priori*. (Kane 2011: 10)

Gemäß Kane ist demzufolge von Gewicht, dass Inferenzen und Behauptungen, die von vornherein als problematisch zu untersuchen gelten, identifiziert werden. Damit die Validierung nicht ein nie endender Prozess wird, können Annahmen, die *a priori* ohne Nachweise als akzeptabel zu betrachten sind und bei denen es keinen Grund zu Zweifeln gibt (vgl. hierzu Kane 2013: 13–15), akzeptiert werden.

Das argumentbasierte Modell von Kane ist u. a. von Chapelle und ihren Kolleginnen (vgl. Chapelle et al. 2008) zur Validierung des TOEFL (*Test of English as a Foreign Language*) angewendet worden und wurde dabei mit zusätzlichen Schritten formalisiert und erweitert. In ihrem Treppenmodell wurden zusätzlich zu den oben erwähnten Inferenzen von Kane beispielsweise auch eine Beschreibung des Bezugsbereichs bei einem Test, eine sog. Domänenbeschreibung (*domain description*), eine Begründung des zu messenden Konstrukts (*explanation*) sowie Belege dafür, dass die Testergebnisse für den beabsichtigten Zweck nützlich sind (*utilization*) hinzugefügt (ibid: 349). Gegenüber einer Orientierung an anderen möglichen Alternativen fanden die Autoren, dass der argumentbasierte Ansatz viele Vorteile angeboten habe (vgl. Chapelle et al.

2010). Die Schritte dieses Treppenmodells sind seitdem in späteren Arbeiten von Chapelle und ihren Kolleginnen weiter verwendet, strukturiert und verdeutlicht worden (vgl. Chapelle et al. 2010; Chapelle & Voss 2013; Knoch & Chapelle 2018; Chapelle 2020).

Auch wenn argumentbasierte Modelle von Praktiken genutzt werden sollten, wurde die Nutzung dieses Ansatzes für eine Bewertung im schulischen Kontext in Frage gestellt (vgl. Moss et al. 2006). Darüber hinaus gebe es zudem Bedenken, da nicht klar definiert sei, wie viele Nachweise gebraucht werden und wie stark diese Argumente sein müssen, um eine bestimmte Inferenz unterstützen zu können (vgl. Newton & Shaw 2014; Xi & Davis 2016).

3.2.2.2 Das soziokognitive Rahmenmodell

Das soziokognitive Rahmenmodell (Weir 2005) gestaltet den Bewertungsprozess, häufig beginnend mit den Testteilnehmenden und ihren Eigenschaften über Testdesign, Bewertung, Testergebnis zu Vergleichen mit externen Kriterien und schließlich den Konsequenzen. Im Rahmenmodell werden, wie der Name bereits andeutet, sowohl die Verwendung der Sprache in einem sozialen Kontext als auch die kognitiven Fähigkeiten und Prozesse der Testteilnehmenden berücksichtigt:

> language use – and language assessment – is both a socially situated and a cognitively processed phenomenon. […] The socio-cognitive framework thus seeks to marry the individual psycholinguistic perspective with the individual and group sociolinguistic perspectives. It could be argued that the socio-cognitive approach helps promote a more "person-oriented" than "instrument-oriented" view of the testing/assessment process than earlier models/frameworks (Shaw & Weir 2007: xi).

Der Fokus liegt somit gemäß den Autoren eher auf den Testteilnehmenden, die im Zentrum des Bewertungsprozesses stehen, als auf dem faktischen Test oder auf den Testinstrumenten. Das soziokognitive Rahmenmodell beschreibt die Schritte zur Testentwicklung und zur Validierung, wobei der Einfluss verschiedener Komponenten aufeinander und die Interaktion zwischen ihnen dargestellt werden, sowohl chronologisch als auch konzeptuell. Auch Weir baut auf das einheitliche Validitätskonzept von Messick auf und unterscheidet dabei die Validität in verschiedene Validitätsarten: Kontextvalidität (*context validity*), kognitive Validität[61] (*cognitive validity*), Validität der

61 In Weir (2005) als theoriebasierte Validität (*theory-based validity*) aufgeführt, in späteren Publikationen aber als kognitive Validität (*cognitive validity*) bezeichnet (vgl. Shaw & Weir 2007).

Ergebnisermittlung (*scoring validity*), kriterienbezogene Validität (*criterion-related validity*) und Konsequenzvalidität (*consequential validity*). Wie diese verschiedenen Arten von Validität miteinander in Verbindung stehen, ist in Abb. 6 dargestellt:

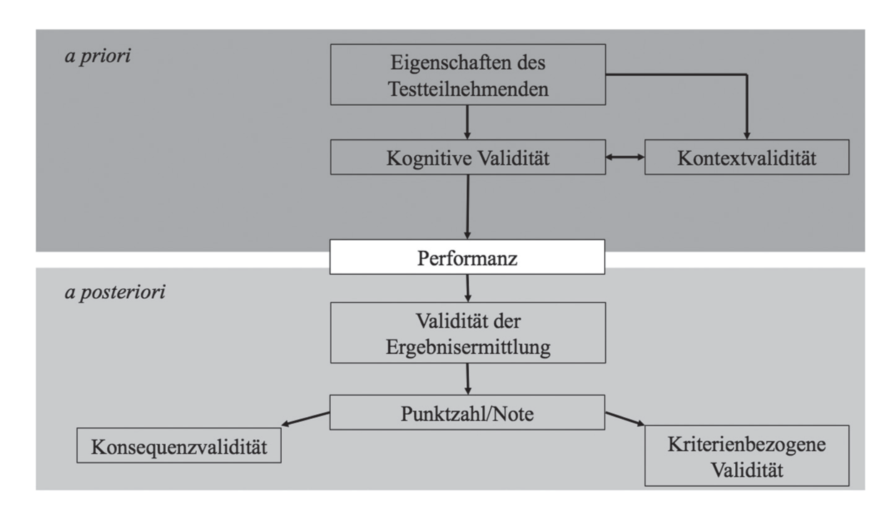


Abb. 6: Darstellung der Hauptkomponenten des soziokognitiven Rahmenmodells zur Testentwicklung und Testvalidierung nach Weir (2005: 47)

Wie in der Abbildung ersichtlich, umfasst das Modell verschiedene Komponenten, wobei die kognitive Validität und die Kontextvalidität generell vor dem Testereignis, *a priori*, die Validität der Ergebnisermittlung, die kriterienbezogene Validität und die Konsequenzvalidität dagegen nach dem Testereignis, *a posteriori*, zu berücksichtigen sind.

Unter *kognitiver Validität* versteht Weir kognitive Prozesse bei der Verarbeitung der Aufgabe, wofür die Testteilnehmenden sowohl sprachliches als auch inhaltliches Wissen benötigen. *Kontextvalidität* bezieht sich auf das Testdesign und die Testdurchführung, traditionell als Inhaltsvalidität (*content validity*) bezeichnet. Gemäß Weir ist die Komponente unter der Bezeichnung Kontextvalidität dargestellt, um auch die soziale Dimension der Sprachverwendung miteinzubeziehen. Kontextvalidität befasst sich zudem u. a. mit der Frage der Abdeckung unterschiedlicher Anforderungen in der Aufgabe und der Repräsentativität einer Aufgabe im Hinblick auf „the larger universe of tasks of which the test is assumed to be a sample“ (Weir 2005: 19). Ein Test bezieht sich somit sowohl auf kognitive Fertigkeiten, die für bestimmte Sprachaktivitäten

gebraucht werden, als auch auf den Kontext, in dem diese Fertigkeiten ausgeübt werden. Die Interaktion zwischen diesen beiden Aspekten, der kognitiven Validität und der Kontextvalidität, und den Bewertungskriterien ist nach Weir „at the heart of construct validity" (2005: 85), wobei er im Gegensatz zu Messick (1989b) eher den Begriff *Validität* als übergeordnete Bezeichnung für das Konzept verwendet (2005: 14).

Gemäß Weir (Weir 2005; Shaw & Weir 2007: 6) ist die *Validität der Ergebnisermittlung* sowohl mit der Kontextvalidität als auch mit der kognitiven Validität verknüpft. *Validität der Ergebnisermittlung* bezieht sich auf die Konsistenz der Testergebnisse, d. h. inwieweit die Testergebnisse über die Zeit stabil sind, inwieweit die Ergebnisse von Ergebnisverzerrung (*bias*) beeinflusst sind und zu welchem Grad die Bewertungskriterien dem zu beurteilenden Konstrukt angemessen sind. Weir (2005) verzichtet hierbei nicht auf die Anwendung der Bezeichnung *Reliabilität*: Nach Weir (ibid.: 14) ist stattdessen von Bedeutung, dass die vorherige Dichotomie von Reliabilität und Validität aufgehoben wird. Die Reliabilität eines Tests sollte als eine eigene Form von Nachweisen hinsichtlich der Validität eines Testverlaufs betrachtet werden und unter dem eigenen Begriff *Validität der Ergebnisermittlung* sollte auf die Reliabilität innerhalb des einheitlichen Validitätskonzepts somit ein erhöhter Fokus gelegt werden (ibid.).

Kriterienbezogene Validität greift auf die traditionelle Definition dieses Begriffs zurück (vgl. Kap. 3.1 oben), wonach das Testergebnis häufig mit einem geeigneten Kriterium oder mit einem älteren etablierten Test der Leistungsfähigkeit korreliert werden kann (*Übereinstimmungsvalidität*). Hierbei kann auch evaluiert werden, inwieweit z. B. die Ergebnisse eines Sprachtests andere Einschätzungen sprachlicher Kompetenz voraussagen können (*Vorhersagevalidität*). Abschließend beinhaltet das Modell auch, inspiriert durch Messick (1989b), die sozialen Konsequenzen einer Bewertung. Zur *Konsequenzvalidität* gehören demnach die Auswirkungen der Bewertung auf die Testteilnehmenden, den Unterricht und die Gesellschaft (sog. *washback effects*). Überlegungen hierzu können nach der erarbeiteten Version sowohl vor dem Testereignis, *a priori*, als auch nach dem Testereignis, *a posteriori*, angestellt werden (vgl. O'Sullivan & Weir 2011). Das soziokognitive Rahmenmodell ist von O'Sullivan und Weir (2011) weiter ausgearbeitet und u. a. auch für Validierungsstudien von Sprachtests verwendet worden (z. B. Wu 2011; Kantarcioglu 2012; Borger 2018). Das Modell verzichtet aber im Kontrast zu den argumentbasierten Rahmenmodellen auf eine systematisch dargestellte Argumentationsstruktur und gibt darüber hinaus wenige Richtlinien hinsichtlich priorisierten Nachweisen (vgl. Xi & Davis 2016).

3.3 Reliabilität und Urteilstendenzen

Die Reliabilität (*reliability*), auch in den *Standards* (American Educational Research Association et al. 2014) in einem eigenen Kapitel dargestellt, zeigt den Grad der Genauigkeit bei Messungen an. Dies bedeutet beispielsweise inwieweit ein Lernender in zwei Prüfungen die gleichen oder ähnliche Ergebnisse erzielt, die durch zwei Bewertende korrigiert werden. Aus dem Beispiel ersichtlich kann sich die Reliabilität somit sowohl auf die Stabilität der Performanz von Lernenden als auch auf die Übereinstimmung zwischen Bewertenden beziehen (z. B. Gipps 1994), wobei der letztere Aspekt in der vorliegenden Studie fokussiert wird. Die Reliabilität zwischen Bewertenden erfährt innerhalb verschiedener Validitätstheorien unterschiedliche Gewichtung. Weir (2005) findet einerseits alle Nachweise zur Validität wichtig und betrachtet somit Untersuchungen zur Reliabilität als einen wichtigen Aspekt der Validität. Gemäß Kane (2013) sind andererseits Diskussionen zur Reliabilität im Hinblick auf die Validität nicht immer relevant, auch wenn einige Nachweise bezüglich der Generalisierung häufig angebracht sind. Des Weiteren wird vorgebracht, dass das Definieren von unterschiedlichen Validitätsaspekten gegen Messicks einheitliches Validitätskonzept laufen würde (vgl. Knoch & Chapelle 2018).

Das Ergebnis eines Tests erklärt, wie gut ein Testteilnehmender eine bestimmte Aufgabe bewältigt hat. Bei der Bewertung von sog. *Performanztests*,[62] z. B. in Form einer Textproduktion, ist die Subjektivität im Bewertungsprozess oft deutlich höher als bei der Bewertung von einem Test des Hörverstehens. Leistungen dieser Art, z. B. Schülertexte, werden vorwiegend von menschlichen Bewertenden – im schulischen Kontext von den Lehrkräften – anhand unterschiedlicher Kriterien bewertet. Dieses performanzbasierte Bewertungsverfahren wird daher auch beurteilergestützte Leistungsmessung genannt.

Das Grundproblem bei Leistungsmessung durch menschliche Beurteiler liegt darin, dass die Leistungen auf der Basis einer subjektiven Bewertung eingestuft werden (z. B. Bachman et al. 1995; Eckes 2011). Daher entsteht die Eventualität, dass die Bewertungen durch zwei voneinander unabhängige Bewertenden unterschiedlich ausfallen. Das Ergebnis einer beurteilergestützten Bewertung kann u. a. von Faktoren wie Merkmalen der Bewertenden (z. B. Alter, Geschlecht, Berufserfahrung und Muttersprache), Merkmalen der Testteilnehmenden (z. B. Alter, Schreibkompetenz in der Erstsprache, Sprachliches

62 Als Performanztest bezeichnet man ein Testverfahren, worin die authentische Verwendung der Sprache durch die Testteilnehmenden in einem handlungsorientierten Kontext evaluiert wird (vgl. Bachman 1990: 304–305).

Niveau und allgemeines Weltwissen) und Merkmalen der Aufgabe und der Testsituation (z. B. Gestaltung, Testdurchführung und informeller/formeller Test) beeinflusst werden. Des Weiteren können zudem Merkmale des Bewertungsrasters oder der Bewertungskriterien (z. B. Grad an Eindeutigkeit) sowie unterschiedliche Bewertungsverfahren (z. B. ein holistisches bzw. analytisches Bewertungsverfahren[63]) einen Einfluss auf die Bewertung haben (z. B. McNamara 1996; Eckes 2005). Zu beachten ist jedoch, dass diese Faktoren eher diffus oder indirekt einen Einfluss auf die Bewertung haben und dass sie auch in Wechselwirkung miteinander treten können (z. B. Eckes 2005).

Für Testentwickler und praktizierende Lehrkräfte ist es wichtig, sich unterschiedlicher möglicher Ursachen mangelnder Beurteilerkonsistenz bewusst zu sein, um diese minimieren zu können. Es ist schwer zu vermeiden, dass unterschiedliche Faktoren einzelner Bewertender die Beurteilung beeinflussen,

63 Um Lernproduktionen fremdsprachlicher Schreibkompetenz zu bewerten, werden häufig Kriterien oder Bewertungsraster, die verschiedene Niveaus von Textqualität beschreiben, verwendet. Diese sind beim Bewertungsprozess als Unterstützung für die Bewertenden gedacht, können aber unterschiedlich gestaltet werden. Bei einem *holistischen* Bewertungsverfahren (eine ganzheitliche Bewertung) wird einer Leistung ein Gesamtergebnis zugeteilt, während unterschiedliche Dimensionen der Leistung durch ein *analytisches* Verfahren (eine Bewertung, bei der bestimmte Kriterien ausgewertet und in ein Gesamtergebnis umgewandelt werden) getrennt bewertet werden. Generell sind sowohl Stärken als auch Schwächen der beiden Bewertungsverfahren identifiziert worden und diese sind weitgehend diskutiert (vgl. McNamara 1996; Weigle 2002; Eckes et al. 2016). Zu den Vorteilen eines holistischen Bewertungsverfahrens gehört eine Emphase in Bezug auf den ganzen Text und nicht die einzelnen Bestandteile. Dazu gilt eine holistische Bewertung als weniger zeitaufwendig. Allerdings kann ein holistisches Verfahren Unterschiede zwischen Bewertenden im Hinblick auf die Interpretation der Kriterien verbergen (vgl. Harsch & Martin 2013). Ein analytisches Verfahren ermöglicht im Gegensatz dazu in größerem Maß einen genaueren Blick auf einzelne Merkmale. Zu den Nachteilen eines analytischen Bewertungsverfahrens gehören, dass es häufig als zeitaufwändig gilt. Crooks und Kollegen (1996: 272) sehen eine Gefahr der beeinträchtigten Validität, wenn eine Bewertung zu analytisch bzw. zu holistisch wird. Bei einer zu analytischen Bewertung besteht die Gefahr darin, dass die Bewertenden ausschließlich die angegebenen Aspekte evaluieren und die globale Perspektive der Leistung übersehen. Vor allem wird aber bei einem analytischen Verfahren häufig befürchtet, dass die einzelnen Bestandteile eine größere Bedeutung als der gesamte Text bekommen werden oder dass einem Aspekt ein allzu starkes Gewicht gegeben wird. Die Gefahr einer zu holistischen Bewertung ist es, dass womöglich eine Note vergeben wird, ohne dass ein breiteres Spektrum von Aspekten beachtet wurde und dass dabei die Stärke bzw. die Schwäche der Leistung nicht hinreichend beachtet wird.

z. B. das Verständnis und die Interpretation der Bewertungskriterien, was damit die Qualität einer beurteilergestützten Bewertung beeinträchtigen kann. Diese sog. Bewertereffekte können unterschiedlicher Art sein. Generell sind folgende Typen von Bewertereffekten in der Forschung untersucht worden: a) *Tendenz zur Strenge bzw. Milde*, b) *Zentraltendenz* (Tendenz zur Mitte), c) *Halo-Effekt* und d) *Primary-Recency-Effekt* (vgl. Bortz & Döring 2002). Tendenz zur Strenge oder Milde bezieht sich auf die Neigung, Leistungen im Vergleich zu anderen Bewertenden tendenziell entweder höher oder niedriger einzustufen. Hierbei kann auch eine sog. differenzielle Strenge bzw. Milde auftreten, wobei der Strenge- oder Milde-Effekt systematisch gewisse Gruppen von Lernenden betrifft oder unter gewissen Umständen hervortritt (vgl. Eckes 2005).

Eine Zentraltendenz liegt dahingegen vor, wenn Bewertende Leistungen hauptsächlich in der Mitte der Ratingskala einstufen. Eine Tendenz zur Mitte kann dann vorkommen, wenn die Bewertenden mit den zu beurteilenden Leistungen wenig vertraut sind oder wenn die Bewerterskala Extrembeispiele nicht berücksichtigt (vgl. Bortz & Döring 2002). Halo-Effekte bezeichnen u. a. die Tendenz, dass ein positives oder negatives Merkmal andere Merkmale oder die Gesamtbewertung überstrahlt. Halo-Effekte kommen häufiger vor, wenn die zu bewertenden Aspekte schwer zu finden und nicht klar definiert sind oder wenn ein Urteil zu schnell gefällt wurde (ibid.).

Bewertereffekte, die als Primary-Recency-Effekte bezeichnet werden können, sind ein Gedächtnisphänomen: bei großer Informationsmenge prägen wir uns die zu Beginn und die zuletzt dargestellte Information bevorzugt ein. Somit können Aspekte, die bei der Bewertung einer Leistung am Anfang oder am Ende herangezogen werden, für das Gesamturteil entscheidend werden. Ein weiteres Beispiel betrifft die Reihenfolge der zu bewertenden Leistungen: eine Leistung mit extremen Merkmalen, die am Anfang beurteilt wird, kann die nachfolgenden Bewertungen beeinflussen (ibid.).

Um den Grad der Bewerterübereinstimmung untersuchen zu können, wird gemäß der *Klassischen Testtheorie* (*Classical test theory, CTT*) traditionell ein Reliabilitätskoeffizient berechnet (z. B. American Educational Research Association et al. 2014) und hierfür wird eine Vielzahl statistischer Methoden verwendet. Diese statistischen Berechnungen haben zum Teil unterschiedliche Eigenschaften, was dazu führen kann, dass sie bei demselben Datenset sowohl eine hohe als auch eine niedrigere Reliabilität aufweisen können (z. B. Eckes 2011). Die Reliabilität ist innerhalb der klassischen Testtheorie das zentrale Konzept. Darüber hinaus kann die Bewerterübereinstimmung innerhalb der *Probabilitistischen Testtheorie* aber auch mittels IRT-Methoden (*Item Response Theory*), insbesondere durch sog. *Multifacetten-Rasch-Modelle,* ausgewertet

werden. Diese beziehen sich nicht nur auf die jeweiligen Bewertenden, sondern auch darauf, wie die Sprachfähigkeit der Lernenden und der Schwierigkeitsgrad der jeweiligen Aufgaben im Verhältnis zueinander stehen (vgl. Hambleton et al. 1991).

Die Reliabilität sollte daher nicht als ein einheitliches Konzept aufgefasst werden, was unpräzise und möglicherweise irreführend sein könnte (Stemler 2004). Stemler weist auf drei Kategorien von Interraterreliabilität hin:

1. Konsensmethoden (*consensus estimates*)
2. Konsistenzmethoden (*consistency estimates*)
3. Methoden zur Messwerteinschätzung (*measurement estimates*)

Konsensmethoden ermitteln den Grad einer genauen Übereinstimmung, wenn unabhängige Bewertende eine Leistung bewerten, während Konsistenzmethoden auf den Grad fokussieren, in dem die bewerteten Leistungen in Relation zueinander stehen, d. h. die relative Reihenfolge der beurteilten Leistungen. Auch wenn Bewertende für Leistungen nicht die gleichen Noten vergeben, was auf einen niedrigeren Konsens der Bewertenden deutet, kann die Reihenfolge der bewerteten Leistungen gleich oder ähnlich sein, was wiederum auf eine hohe Konsistenz der Bewertenden hindeutet.

Methoden zur Messwerteinschätzung, die dritte Art von Interraterreliabilität, werden oft mit einem sog. Multifacetten-Rasch-Modell ermittelt (vgl. Eckes 2015; 2019). Durch diesen Ansatz können unterschiedliche Informationen, sog. Facetten wie z. B. der Grad der Strenge bzw. Milde der Bewertenden oder der Schwierigkeitsgrad unterschiedlicher Aufgaben, eingeschätzt werden. Zu beachten ist aber, dass hohe Reliabilitätswerte nicht notwendigerweise bedeuten, dass die Prüfung oder die Interpretation der Prüfungsergebnisse auch valide ist. Bewertende können eine hohe Übereinstimmung aufweisen und dennoch nicht die zu testenden Kompetenzen in Betracht ziehen (vgl. Koretz 2008). Bei der Auswahl von Methoden zur Reliabilitätsbestimmung sollte außerdem sorgfältig überlegt werden, inwiefern die Methoden mit Blick auf die Eigenschaften der Daten adäquat und angemessen sind (vgl. hierzu die *Standards*, American Educational Research Association et al. 2014), und nach Eckes, Müller-Karabil und Zimmermann (2016) sollte hierbei immer mindestens ein Konsens- und ein Konsistenzwert berechnet werden.

3.4 Fazit

Der konzeptuelle Rahmen der vorliegenden Arbeit sollte zum Verständnis der untersuchten Phänomene und gleichzeitig zur Beantwortung der

Forschungsfragen beitragen. Wie hier gezeigt, haben Definitionen, Konzepte und Modelle zur kommunikativen Kompetenz einen deutlichen Einfluss auf heutige Bildungsstandards im schwedischen Schulkontext sowie auf den GER, den Referenzpunkt für das gegenwärtige schwedische System. Damit kann angenommen werden, dass Bewertungen, die Bewertungsskalen dieser beiden Ansätze verwenden, auch Elemente der handlungsorientierten Betrachtungsweise enthalten. Die Kompetenzmodellierungen können als Grundlage für ein besseres Verständnis davon dienen, wie die Bewertungen das zu bewertende Konstrukt wiederspiegeln und wie die Bewertenden in ihren Kommentaren ihr Verständnis für das zu messende Konstrukt konzeptualisieren.

Dieses Kapitel beinhaltet zudem eine Beschreibung der in der vorliegenden Arbeit zentralen Qualitätsindikatoren hinsichtlich Interpretation und Verwendung der Ergebnisse von Sprachtests. Das Konzept der Validität hat sich seit Mitte des 20. Jahrhunderts wesentlich verändert. Von einer dreigliedrigen Einteilung ging die Entwicklung hin zu einer einheitlicheren Definition der Validität mit verschiedenen Facetten, einschließlich Dimensionen wie soziale Werte und Konsequenzen (vgl. Messick 1989a). In neueren Definitionen liegt der Fokus somit nicht, wie zuvor, auf dem faktischen Test, sondern eher auf der Interpretation und Verwendung der Testergebnisse. Zusammenfassend lässt sich ableiten, dass das Validitätskonzept als vielseitig zu betrachten ist, aber auch, dass es bis heute in der Forschung keineswegs Konsens zu verschiedenen Aspekten der Validität und deren Anwendung gibt. Dies zeigt sich nicht zuletzt, wenn man den Umgang mit dem Begriff Konstruktvalidität betrachtet. Während Messick die Konstruktvalidität als das übergeordnete Konzept der Validität betrachtet, wird dem Begriff im sozio-kognitiven Modell von Weir (2005) explizit kein Platz gegeben. Das einheitliche und erweiterte Validitätskonzept von Messick hat sich für Validierungsstudien auch angesichts eines schwedischen Schulkontexts als angemessen erwiesen (z. B. Nyström 2004: Klapp-Lekholm 2008) und gilt immer noch als wichtig und relevant für gegenwärtige Untersuchungen der Validität im Bereich des Testens (vgl. Chapelle 2020). Dies zeigt sich nicht zuletzt auch im Hinblick darauf, dass viele seiner Nachfolger auf das einheitliche Validitätskonzept aufgebaut haben und seine Ideen beispielsweise in den Rahmenmodellen zur Validierung von Weir (vgl. 2005) und Kane (vgl. 2006; 2013) aufgegriffen werden. Dazu zeigen aber die verschiedenen Definitionen und die uneinheitliche Entwicklung des Konzeptes deutlich die Komplexität der Lage in Fragen der Validität, was auch zu unterschiedlichen Modellen und Praktiken hinsichtlich Validierungen von Tests und Testverwendung führt.

Im Mittelpunkt dieser Studie steht eine empirisch basierte Validitätsuntersuchung. Theoretische Ansätze zur Erfassung verschiedener Validitätsaspekte werden hauptsächlich vom argumentbasierten Ansatz nach Kane (2006; 2013) und dem soziokognitiven Rahmenmodell gemäß Weir (2005) bezogen. Das Ziel der vorliegenden Arbeit ist jedoch nicht, eine vollständige Validierung eines Tests oder eines Bewertungsverlaufs durchzuführen und aus diesem Grund wird auf einen rein argumentbasierten Ansatz verzichtet. Da der Fokus der vorliegenden Studie darin liegt, unterschiedliche Validitätsaspekte bei einer Bewertung schriftlicher Sprachkompetenz zu untersuchen, wird zudem nicht das gesamte soziokognitive Ramenmodell von Weir berücksichtigt. Die Untersuchung bezieht sich somit eher darauf, wie Leistungen schwedischer Lernenden evaluiert werden und wie diese Testergebnisse interpretiert und verwendet werden können. Der Test, der in dieser Studie genutzt wird, sollte bereits im Hinblick auf den Inhalt und die beabsichtigten kognitiven Kompetenzen auf ein B1-Niveau vor dem Testereignis, *a priori*, kalibriert sein. Im Mittelpunkt steht daher die Bewertung der Leistungen, die nach dem Testereignis, *a posteriori*, geschieht. In der vorliegenden Studie können folglich sowohl Kanes Argumentationsstruktur zur Validierung als auch verschiedene Aspekte der Validität bei Messick und Weir bei der Interpretation der Daten nützlich sein, insbesondere Validitätsaspekte, die mit der Bewertung nach dem Testereignis zu tun haben. Da unterschiedliche Aspekte der Validität schwer zu trennen sind und miteinander in enger Beziehung stehen (vgl. Weir 2005), werden auch andere Validitätsaspekte beachtet. Aspekte der Reliabilität sind hierbei ebenfalls ein wichtiger Teil dieser Untersuchung, nicht zuletzt, da ein akzeptables Maß an Reliabilität bei einer Bewertung eine Voraussetzung für die Validität ist und da eine hohe Reliabilität ein Zeichen dafür sein könnte, dass Bewertende ein gemeinsames Verständnis für das zu messende Konstrukt haben.

Eine Studie zur Validierung bedarf Nachweisen unterschiedlicher Art und beschäftigt sich somit häufig sowohl mit quantitativen als auch mit qualitativen Methoden. Eine Kombination quantitativer und qualitativer Forschungsmethoden, ein sog. Mixed-Methods-Ansatz also, hat sich in vielen Validierungsstudien als eine vorteilhafte und nützliche Perspektive erwiesen (vgl. Borger 2018; Chapelle 2020) und ein Ansatz jener Art wird demzufolge auch in der vorliegenden Arbeit verfolgt (vgl. Kap. 5).

4. Stand der Forschung

Dieses Kapitel verortet die im Hinblick auf die vorliegende Studie relevante bisherige Forschung zur Bewertung von Sprachkompetenzen in einer Fremdsprache. Hierfür soll an erster Stelle auf Studien, die verschiedene Validitätsaspekte bei der Bewertung von L2-Lernerproduktionen fokussiert haben, eingegangen werden. Von Relevanz für diese Untersuchung sind insbesondere empirische Studien zur Bewertung fremdsprachlicher Kompetenz aus einem schwedischen Schulkontext und wissenschaftliche Arbeiten, die eine Bewertung fremdsprachlicher Schreibkompetenz ausgewertet haben. Empirische Forschungsarbeiten zur Bewertung fremdsprachlicher Kompetenz finden sich aber in einem schwedischen Kontext, trotz eines erhöhten Interesses für Bildungsstudien und Projekte, unabhängig von Schwerpunkt nur in geringerer Anzahl, vor allem hinsichtlich der zweiten Fremdsprache (vgl. Bardel et al. 2019).

Es handelt sich bei den bisher durchgeführten empirischen Untersuchungen überwiegend um Untersuchungen in der ersten Fremdsprache Englisch (z. B. Erickson 2009; Borger 2018). Bislang wurde außerdem auch Aufmerksamkeit auf die Beschreibung und Ermittlung der Beurteilerkonsistenz gerichtet. Dazu zählen Studien, die darauf fokussieren, inwieweit Lehrkräfte untereinander eine ausreichende Beurteilerübereinstimmung aufweisen, was bisher zum Teil divergierende Forschungsbefunde erzeugt hat (vgl. Skolverket 2009; Skolinspektionen 2017). Generell scheint die Bewerterübereinstimmung in einem schwedischen Schulkontext verhältnismäßig viel Beachtung erhalten zu haben, insbesondere nach den in den Medien oft diskutierten Kontrollkorrekturen der schwedischen Schulaufsichtsbehörde von 2010 bis 2019 (z. B. Skolinspektionen 2010; 2018) sowie Untersuchungen zum Verhältnis zwischen einem Ergebnis der nationalen Prüfung und der Abschlussnote (z. B. Skolverket 2020a). Insgesamt gibt es dagegen nur wenige empirische Arbeiten, die sich mit dem Bezug zum GER aus einer schwedischen Perspektive befasst haben.

Das Verwenden performanzbasierter Bewertungen, wenn Aussagen über die schriftliche Kompetenz eines Individuums getroffen werden sollen, ist heutzutage eine etablierte Prüfungsform in einer Fremdsprache. Diese verlangen eine beurteilergestützte Evaluation, die notwendigerweise von subjektiver Art ist und somit zu *Variabilität* bei den Bewertungen führen kann. Variabilität kann in Bewertungen anhand unterschiedlicher Eigenschaften der Bewertenden betreffender Faktoren wie Erfahrungen, Alter, Muttersprachler vs. Nicht-Muttersprachler, Interpretation und Verständnis der Bewertungsskala und

Strenge-Milde-Tendenz der Bewertenden, sog. *Bewertereffekten* reflektiert werden (vgl. Cumming 1990; Eckes 2008; Zhang & Elder 2011). Auch die Auswirkungen u. a. unterschiedlicher Bewertungsverfahren auf den Bewertungsprozess oder der Effekt von Bewertertraining standen in den letzten Jahrzehnten in mehreren Studien im Fokus, um ein tieferes Verständnis für den faktischen Bewertungsprozess zu erlangen (vgl. Weigle 1994; Lumley 2002; Barkaoui 2011a). Im Bereich des Fremdsprachentestens wird der GER als Bezugsrahmen immer häufiger verwendet, nicht zuletzt für die Anbindung von Ergebnissen standardisierter Tests an die unterschiedlichen Referenzniveaus (z. B. Papageorgiou 2007; O'Sullivan 2008; Kecker 2011; Tschirner & Bärenfänger 2012; Papageorgiou et al. 2015; Green 2018; North & Piccardo 2018). Zunehmend haben aber auch Bezugsstudien hinsichtlich fremdsprachlicher Kompetenzen von Lernenden[64] (z. B. European Commission 2012b; Abel et al. 2012; Bärenfänger 2016; Aronsson 2020) den Referenzrahmen benutzt.

Zunächst werden Studien zum Bewerterfokus vorgestellt, d. h. *was* Bewertende bei einer Beurteilung von Leistungen berücksichtigen (Kap. 4.1). Im Anschluss werden Studien bezüglich *Beurteilerübereinstimmung* präsentiert, d. h. in welchem Ausmaß Einstufungen durch unterschiedliche Bewertende zu *ähnlichen* oder *denselben* Testergebnissen kommen (Kap. 4.2). Danach wird die Forschungslage relevanter Sprachleistungs- und Validierungsstudien, die den GER als Bezugspunkt nehmen, vorgestellt (Kap. 4.3). Abschließend werden die Befunde im Hinblick auf die Relevanz für die vorliegende Arbeit zusammenfassend diskutiert (Kap. 4.4).

4.1 Bewertung fremdsprachlicher Kompetenz – Fokus der Bewertenden

Lernerproduktionen zu beurteilen ist ein komplexer Prozess, wobei Bewertende in der Regel Strategien entwickeln, unterschiedliche Kriterien berücksichtigen und Schlüsse ziehen müssen, um zu einem Urteil zu kommen. Generell hat sich gezeigt, dass Bewertende unter Umständen auf gewisse Aspekte fokussieren und dabei andere vernachlässigen (vgl. McNamara 1990; Cumming et al. 2002; Lumley 2002). Welche Aspekte und Bewertungsdimensionen Bewertende

64 Viele bisherige Untersuchungen zur Evaluation fremdsprachlicher Teilkompetenzen in Bezug auf den GER beachteten hauptsächlich Leistungen in den rezeptiven Kompetenzbereichen Hören und Lesen (vgl. Köller et al. 2010) und verzichteten dabei auf die produktiven Fertigkeiten, vor allem die mündliche Kompetenz, die bei der Überprüfung als aufwendiger zu erheben und zu bewerten galten.

berücksichtigen und gewichten, kann folglich zwischen einzelnen Bewertenden oder verschiedene Bewertergruppen mitunter stark variieren (vgl. Politt & Murray 1996; Eckes 2008; Kim 2009; Borger 2018), was auch dazu führen kann, dass Bewertungsurteile unterschiedlich ausfallen. Bewertervariabilität kann somit ihren Grund in unterschiedlichen Interpretationen des zu messenden Konstruktes durch die Bewertenden (vgl. *Konstruktvalidität*, Kap. 3.2) haben.

Dieses Kapitel widmet sich zunächst der Forschungslage und Ergebnissen für die vorliegende Arbeit relevanter Studien im Hinblick auf die *Konstruktkonzeptualisierung* von Bewertenden. Eine Bewertervariabilität hinsichtlich der Konstruktkonzeptualisierung zeigt sich u. a. darin, dass nicht alle Urteilsdimensionen bewertet werden, dass der erste Leistungseindruck bleibt oder dass ein hervortretender Aspekt weniger auffallende Bewertungsdimensionen beeinflusst (vgl. Lumley 2002; Eckes 2008). Bisherige Studien zur Konstruktkonzeptualisierung von Bewertenden[65] sind zum Teil zu divergierenden Ergebnissen im Hinblick auf zentral beachtete Aspekte bei einer Bewertung gekommen. Dies hat womöglich u. a. mit unterschiedlichen Kontexten, Beurteilungsverfahren sowie verschiedenen Bewerter- und Lernergruppen zu tun. Studien zur Konstruktkonzeptualisierung von Bewertenden fokussieren häufig entweder auf Bewertercharakteristiken oder auf Merkmale im Bewertungsprozess. Die Variabilität bei einer Bewertung fremdsprachlicher Leistungen ist häufig auf unterschiedliche Konzeptionen und Verhalten von Bewertenden zurückzuführen. Eine Bewertervariabilität weist damit eher auf Umstände, Kontexte oder Eigenschaften der Bewertenden hin (vgl. McNamara 1990; Eckes 2008) und hat oft weniger mit der Leistung des einzelnen Lernenden zu tun, obwohl auch Leistungsmerkmale der Lernerproduktionen zu Unterschieden führen können (vgl. Pollitt & Murray 1996).

Die überwiegende Mehrheit bisheriger Untersuchungen hat Schwerpunktsetzungen bei der Bewertung von Lernerleistungen in Englisch als Fremd- oder

65 Für die vorliegende Arbeit sei nicht nur wissenschaftliche Arbeiten zur Schwerpunktsetzung auf schriftliche Kompetenz (vgl. Lumley 2002; Barkaoui 2010a; 2010b; Kuiken & Vedder 2014), sondern auch auf die vielen Studien hinsichtlich mündlicher Kompetenz hingewiesen (vgl. Magnan 1988; Pollitt & Murray 1996; Brown et al. 2005; Iwashita et al. 2008; Hsieh 2011; May 2011; Bøhn 2016; Borger 2018). Auch wenn Bewertungen mündlicher Kompetenzen zum Teil auch andere Bewertungsdimensionen berücksichtigen, wie z. B. die *Aussprache* oder die *Interaktion*, können Ergebnisse jener Studien auch für Untersuchungen von Bewertungen schriftlicher Kompetenzen relevant sein (vgl. Kuiken & Vedder 2014). Der hauptsächliche Fokus liegt jedoch auf Studien zur Bewertung schriftlicher Kompetenzen.

Zweitsprache vorgenommen und stammt aus einem nicht-schwedischen Schulkontext. Die Befunde können trotzdem für die vorliegende Arbeit von Relevanz sein. Bisherige Studien zeigen generell, dass Bewertende mehrere unterschiedliche Aspekte bei der Bewertung fremdsprachlicher Leistungen beachten. Es handelt sich dabei um Aspekte wie *Korrektheit*, *Spektrum*, *Kohärenz*, *Interaktion*, *Flüssigkeit*, *Aufgabenerfüllung, Strategien* oder *Verständlichkeit* (vgl. Brown et al. 2005; Iwashita et al. 2008; Borger 2018).

Wissenschaftliche Arbeiten, die den Bewerterfokus hinsichtlich fremdsprachlicher Kompetenzen ausgewertet haben, liegen innerhalb eines schwedischen Kontextes nur in sehr überschaubarer Anzahl vor: Die für die vorliegende Arbeit relevante Studie von Borger (2018) ist in diesem Zusammenhang eine Ausnahme. Zusammenfassend hält Borger in ihrer Untersuchung der Konstruktkonzeptualisierung bei der Bewertung mündlicher Leistungen in der ersten Fremdsprache Englisch fest, dass schwedische Bewertende ein breites Spektrum unterschiedlicher Aspekte bei der Bewertung beachten. Davon sind Kommentare zur *Korrektheit* die üblichsten im Material. Dies deutet darauf hin, dass Korrektheit eine tragende Rolle für schwedische Bewertende im Bewertungsprozess hat. Weitere schwerwiegende Aspekte sind gemäß der Studie die *Differenziertheit*, die *Kohärenz*, die *Interaktion* und die *Flüssigkeit.* Auch wenn die GER-Bewertenden in der Studie hauptsächlich die gleichen Aspekte wie die schwedischen Bewertenden berücksichtigt, kommen die *Diffenziertheit* und die *Flüssigkeit* bei ihnen vor der *Korrektheit.* Die von den GER-Bewertenden beachteten Aspekte sind zudem gleichmäßiger auf die meistbeachteten Bewertungsdimensionen verteilt. Weniger Beachtung in dieser Untersuchung erhalten Aspekte wie *Verständlichkeit*, *Aufgabenerfüllung*, *Strategien* und *soziolinguistische Kompetenz.* Auch weitere Studien haben gezeigt, dass Bewertende in ihren Beurteilungen gelegentlich mehr Wert auf bestimmte Bewertungsdimensionen im Vergleich zu anderen legen. Insbesondere scheinen, wie in der Studie von Borger (2018), Aspekte der linguistischen Kompetenz, wie *Formale Strukturen* und *Wortschatz*, häufiger von Bewertenden berücksichtigt zu werden (z. B. Magnan 1988; McNamara 1990; 1996).

Die für die vorliegende Untersuchung relevanten Erkenntnisse sind zudem, inwiefern die Bewertenden positive bzw. negative Urteile im Hinblick auf die berücksichtigten Bewerteraspekte fällen (vgl. Vaughan 1991; Rinnert & Kobayashi 2001; Barkaoui 2010a; Borger 2018). In vielen dieser Studien kann übereinstimmend eine Tendenz zur strengeren Beurteilung bezüglich der Korrektheit wahrgenommen werden: Bewertungsdimensionen bezüglich *sprachlicher Korrektheit* scheinen somit generell eine negative Bewertung zu erhalten, während Aspekte wie *Kohärenz* und *Flüssigkeit* öfter in positiven Worten beschrieben

werden. Zusammenfassend kann festgestellt werden, dass gerade sprachliche Mittel, insbesondere *sprachliche Korrektheit*, im Vergleich zu weiteren Aspekten wie solchen der *Verständlichkeit* oder *Flüssigkeit*, strenger beurteilt werden (McNamara 1990; 1996; Eckes 2008). Dies deutet ebenfalls auf einen gewissen Fokus auf sprachliche Korrekturen bei der Bewertung fremdsprachlicher Kompetenz. Für einen Fokus auf Sprachkorrekturen und somit eine eher fehlerorientierte Beurteilung kann es mehrere Gründe geben. Birkel und Birkel (2002) weisen auf eine Neigung der Lehrkräfte hin, Kriterien heranzuziehen, die einfach zu erfassen sind, wenn sie Schülertexte bewerten. Sprachliche Korrekturen sind auch objektiv leichter zu begründen als z. B. inhaltliche Anforderungen. Es wird in diesem Zusammenhang u. a. auch angeführt, dass vage formulierte Bewertungskriterien und Deskriptoren ein Grund für einen Fokus auf grammatische Korrektheit sein könnten. Wisniewski (2010) hat ähnliche Schlüsse gezogen: Sie zeigt in ihrer Studie zur Bewertervaliabilität, dass Bewertende häufig auf andere Kriterien, die nicht in den Skalen vorhanden sind, wie z. B. Anzahl der Fehler, zurückgreifen. Das könne u. a. seinen Grund darin haben, dass die Formulierungen der Bewertungsskalen und Kriterien vage seien, was Bewertende mit einer Tendenz zur Überbetonung von Korrektheit und somit einer fehlerorientierten Beurteilung kompensieren würden (ibid.). Auch Aspekte, die nicht explizit in den Skalen vorkommen, scheinen somit bei einer Bewertung eine gewisse Rolle zu spielen. Zu diesen zählen bei der Bewertung schriftlicher Leistungen beispielsweise Aspekte wie die *Textlänge* (Lumley 2002; Barkaoui 2010a, Håkansson Ramberg 2021) oder die *Anzahl der Fehler* (Wisniewski 2010).

Lumley (2002) ist in einer Studie zu Bewerterstrategien bei der Beurteilung von Lernertexten zu dem Ergebnis gekommen, dass Bewertende versuchen, die Bewertungskriterien zu verwenden, jedoch ihre Urteile auf ein komplexes und unbestimmbares Gefühl des Textes unabhängig von Formulierungen in den Bewertungskriterien gründen. Gemäß der Studie stoßen sie in den Texten häufig auf Eventualitäten, die nicht von den Skalen abgedeckt werden. Die Bewertenden fühlen sich jedoch offenbar verpflichtet, die Formulierungen der Kriterien zu verwenden, wobei die Bewertungskriterien eher für das Formulieren einer nachträglichen Rechtfertigung der Beurteilung funktionierten (ibid.).

Die Bedeutung von Bewertercharakteristiken wie Alter, Ausbildung und Erfahrung der Bewertenden wird in mehreren wissenschaftlichen Arbeiten fokussiert. Im Mittelpunkt steht dabei u. a. die Erfahrungen der Bewertenden (vgl. Cumming 1990; Weigle 1994; Eckes 2008). In einer Studie von Cumming (1990) wurden mithilfe von *Think-aloud-protocols* u. a. Strategien von erfahrenen bzw. weniger erfahrenen Bewertenden bei der Bewertung von

Textproduktionen in Englisch als Fremdsprache untersucht. Die Ergebnisse zeigen, dass der Bewertungsprozess erfahrener Bewertender komplexerer schien. Erfahrene Bewertende können effektiv textbasierte und situationsabhängige Strategien gleichzeitig integrieren, wobei die Bewerterurteile erfahrener Bewertenden ein breiteres Spektrum von Aspekten umfassten als die weniger erfahrener Bewertender (ibid.). In anderen Studien konnten zudem unterschiedliche Bewerterprofile, partiell korrelierend mit gewissen Hintergrundvariabeln, identifiziert werden, wie z. B. ein *grammatikorientierter Stil* (Vaughan 1991; Cumming et al. 2002; Eckes 2008). Da Bewertercharakteriska in der vorliegenden Arbeit nicht im Zentrum stehen – ohnehin wurden die Textproduktionen größtenteils von ausgebildeten und erfahrenen Bewertenden beurteilt – werden diese Erkenntnisse nur am Rande der Untersuchung miteinbezogen.

In einer Studie von Kim (2009) wurde die Variabilität bei der Bewertung mündlicher Kompetenz durch englische Muttersprachler bzw. Nicht-Muttersprachler des Englischen ausgewertet. Die Ergebnisse zeigen, dass die Bewertenden hauptsächlich Aspekte zur *Aussprache* und zum *Wortschatz* berücksichtigt hatten. Die Begründungen der Muttersprachler waren jedoch ausführlicher, in größerem Ausmaß auf die verschiedenen Bewertungskriterien verteilt und gaben detailliertere Beschreibungen der Bewertungskriterien. Die Nicht-Muttersprachler waren hingegen allgemeiner in ihren Bewerterkommentaren.

Des Weiteren kann auf eine Diskrepanz zwischen Bewertenden verwiesen werden, die auf kulturelle Unterschiede zurückzuführen ist: Bewertende, die einen gemeinsamen professionellen und kulturellen Hintergrund teilen, scheinen ähnlichen Erwartungshaltungen zu folgen, auch wenn dies nicht allzu auffällig zu Tage tritt (vgl. Song & Caruso 1996; Cumming et al. 2002). Erfahrene Lehrkräfte eines Faches scheinen zudem eine Vorstellung davon zu haben, was von Schülerinnen und Schülern unterschiedlicher Jahresstufen zu erwarten ist, und ihre Bewertungen nach diesen Maßstäben zu begründen (vgl. Jølle 2015). Dies könnte aber problematisch sein, da Lehrkräfte anscheinend nicht immer mit dem übereinstimmen, was von den Lernenden auf gleichem Niveau zu erwarten ist (vgl. Håkansson Ramberg 2021).

In einer Studie untersuchten Kuiken und Vedder (2014) die Bewertung schriftlicher L2-Produktionen von Lernenden des Niederländischen bzw. Italienischen, die etwa auf den Niveaus A2–B1 des GER anzusiedeln sind. Dabei konnte wahrgenommen werden, dass Bewertende verschiedene Strategien zu verwenden scheinen, wenn sie Texte auf einem höheren bzw. niedrigeren Niveau beurteilen. Bei der Bewertung schriftlicher Leistungen auf niedrigeren

Niveaus scheinen wie auch in einer Studie von Pollitt und Murray (1996) Aspekte wie *Verständlichkeit* von besonderem Gewicht zu sein. Die Ergebnisse dieser Studien deuten darauf, dass auch das generelle Sprachkompetenzniveau der Lernenden von Bedeutung sein kann, wenn die Konstruktkonzeptualisierung von Bewertenden untersucht werden soll.

4.2 Bewerterübereinstimmung bei schriftlichen Leistungen

In Studien zur Variabilität bei einer beurteilergestützten Einschätzung fremdsprachlicher Produktionen steht häufig die Beurteilerkonsistenz im Zentrum. Sie ist ein wichtiger Bestandteil bei validen Rückschlüssen auf die Sprachfertigkeit eines Prüfungsteilnehmenden in der Fremdsprache. Dabei hat sich herausgestellt, dass selbst geschulte und erfahrene Bewertende Kriterien unterschiedlich verwenden und eine Variabilität bezüglich ihrer Bewerterurteile aufweisen (vgl. Eckes 2008). In vielen der bisherigen Studien fanden sich zudem widersprüchliche Befunde u. a. im Hinblick auf den Effekt von Bewertertrainings, Eigenschaften der Bewertenden, z. B. Milde- bzw. Strengetendenzen, oder die Verwendung analytischer bzw. holistischer Bewertungsinstrumente, welche ebenfalls die Beurteilerübereinstimmung beeinflussen können (vgl. Lumley 2002; 2005; Eckes 2011). Immerhin konnten durch Studien in den letzten Jahren, die umfangreiche Daten und zum Teil andere Analysemethoden verwendeten, neue Erkenntnisse gewonnen werden.

Ein Problem scheint im schwedischen Schulkontext vor allem die ungenügende Beurteilerübereinstimmung bei der Benotung schriftlicher Lernerproduktionen darzustellen. Kritik im Hinblick auf die Bewerterübereinstimmung haben vor allem Aufsätze im Fach Schwedisch (d. h. in der Mehrheitssprache) erhalten, in vielen Fällen aufgrund einer negativen Differenz (vgl. Skolinspektionen 2010; 2017). Die Befunde haben zudem gezeigt, dass die Bewertungsergebnisse der schriftlichen Lernerproduktionen für das Schulfach Schwedisch in höherem Ausmaß als die Ergebnisse in der ersten Fremdsprache Englisch abweichen (z. B. Skolinspektionen 2017). Darüber hinaus hat es sich erwiesen, dass die höchste Beurteilerübereinstimmung zwischen Texten mit der niedrigsten Benotung vorliegt (ibid.). Indes sind große Unterschiede zwischen den Schulen zu finden, wobei die Befunde insgesamt darauf hingewiesen haben, dass Lehrkräfte im Vergleich zu externen Bewertenden bessere Noten vergeben (vgl. Skolinspektionen 2018).

Diese von der Schulaufsichtsbehörde wiederholten Korrekturen durch externe Bewertende haben jedoch von den Wissenschaftlern Gustafsson und

Erickson (2013) Kritik erhalten. Sie meinen, dass die externen Bewertenden bei ihrer Beurteilung andere Voraussetzungen hatten als die Lehrkräfte, u. a. verwendeten die externen Bewertenden eine andere Bewertungsskala und sie erhielten hierzu schlecht kopierte Schülerleistungen, was zu einem niedrigeren Urteil geführt haben könnte. Des Weiteren ergibt sich die Frage nach der Auswahl der externen Bewertenden. Ihre Beurteilungen wurden zudem von der schwedischen Schulaufsichtsbehörde ungeachtet ihrer Kompetenz als objektiver Nachweis im Vergleich zu den Bewertungen der praktizierenden Lehrkräfte angesehen (ibid.).

Empirische Forschungsarbeiten haben nachgewiesen, dass Lehrkräfte eine geringfügige Tendenz zur Milde haben können, wenn sie die eigenen Schülerinnen und Schüler bewerten (vgl. Östlund-Stjärnegårdh 2002; McKinstry et al. 2004; Harlen 2005; Hambleton et al. 1995), wobei andere Studien weniger Tendenzen in diese Richtung gezeigt haben (vgl. Birkel & Birkel 2002; Gibbons & Marshall 2010). Mögliche Bewertereffekte können auch für die externen Bewertenden in Frage kommen, u. a. ihre Rolleninterpretation in Bezug auf die Notwendigkeit von Strenge bei ihren Beurteilungen. Die Tatsache, dass die externen Bewertenden in gewissem Maße selbst ausgewählt wurden (sog. *self-selction-bias*) und dementsprechend nicht notwendigerweise als repräsentativ für die Lehrerpopulation in Schweden anzusehen sind, könnte ebenfalls zum abweichenden Ergebnis beigetragen haben (vgl. Gustafsson & Erickson 2013).

Wissenschaftliche Studien im schwedischen Kontext zeigen darüber hinaus vor allem bei längeren Aufsätzen im Fach Schwedisch eine niedrigere Bewerterübereinstimmung, wobei Unterschiede zu der ersten Fremdsprache Englisch bezüglich der Beurteilerkonsistenz bei der Bewertung von Textproduktionen relativ gering sind (vgl. Gustafsson et al. 2014). Ergebnisse einer Studie von Erickson (2009) zeigen eine hohe Beurteilerkonsistenz für schwedische Bewertende bei der Bewertung schriftlicher Leistungen, Rangkorrelationswerte nach Spearman's Rho liegen zwischen .86 und .93. Unter den Bewertenden können zudem unterschiedliche Profile wahrgenommen werden, wie eine Zentraltendenz (d. h. eine Tendenz mittleren Noten zu vergeben) sowie leichte Tendenzen zur Milde bzw. Strenge. In einer weiteren, bereits erwähnten Studie, hat Borger (2018) die Bewertung mündlicher Sprachfertigkeit im Englischen von Lernenden am Gymnasium untersucht. Die Untersuchung von Borger konnte im Hinblick auf die Bewerterübereinstimmung, ähnlich wie in der Studie von Erickson, eine zufriedenstellende Beurteilerkonsistenz, mit Werten zwischen .59 und .95 (Medianwert .77) nach Spearman's Rho und zwischen .47 und .89 (Medianwert .66) nach Kendalls Tau-b, feststellen. Des Weiteren zeigen die schwedischen Bewertenden in der Studie mit dem Ergebnis .98 zudem eine gute

innere Konsistenz (Cronbachs Alpha).[66] Weitere Ermittlungen zur Reliabilität, z. B. bezüglich Konsenswerten, sind in Borgers Studie nicht durchgeführt worden. Andere Untersuchungen verschiedener Art aus dem schwedischen Schulkontext haben über die Jahre zeigen können, dass Bewertungen von Lehrkräften Mängel hinsichtlich der Vergleichbarkeit zwischen Klassen und Schulen aufweisen. Die Lehrkräfte scheinen ihre eigenen Schülerinnen und Schüler in der Klasse in ein Verhältnis zueinander setzen zu können, haben es aber schwerer, die Ergebnisse der eigenen Klasse gegenüber Leistungen anderer Schulklassen einzuschätzen (vgl. SOU 1942:11; Johansson 2013).

Für die vorliegende Arbeit sind nicht nur Untersuchungen fremdsprachlicher Kompetenz, sondern auch Studien zur Bewerterübereinstimmung hinsichtlich schriftlicher Kompetenz aus dem schwedischen Kontext relevant, da es bei der Bewertung freier Produktion Überschneidungen geben könnte. In einer Studie von Dalberg (2019) wurde die Übereinstimmung von Lehrkräften im schwedischen Gymnasium[67] hinsichtlich der Bewertung zweier Tests des schriftlichen Ausdrucks innerhalb der nationalen Prüfung im Fach Schwedisch untersucht. Dabei wurden auch der Generalisierbarkeitskoeffizient für einzelne Bewertungen einer Lehrkraft und für Paarbewertungen zweier miteinander diskutierender Lehrkräfte ermittelt, um Effekte von einzelnen bzw. Paarbewertungen und der Anzahl der Bewertenden festlegen zu können. Die Ergebnisse der Berechnungen zeigen, dass die Generalisierbarkeit mit einer höheren Anzahl von Bewertenden zunahm, aber die Steigerung nach zwei Bewertenden im Wesentlichen abnimmt. Der Unterschied zwischen einzelnen Bewertungen und paarweisen Bewertungen im Hinblick auf die Zuverlässigkeit war jedoch gering. Die Ergebnisse dieser Studie weisen dementsprechend deutlich darauf hin, dass mindestens zwei Bewertende jede Schülerleistung beurteilen sollten – der weitere Gewinn durch zusätzliche Bewertenden wird mit der Anzahl

66 In Kap. 5.3.3 werden Methoden zur Bestimmung der Bewerterübereinstimmung ausführlicher erläutert.

67 Zu bemerken ist, dass die Lehrkräfte dieser Studie nicht durch Zufall ausgesucht wurden, sondern zu Referenzgruppen für die nationalen Prüfungen in Schwedisch gehören, die sich regelmäßig treffen, um Beurteilungen von Schülerleistungen und Prüfungsanweisungen zu diskutieren. Hierbei kann angenommen werden, dass diese selbstselektierte Gruppe von Lehrkräften ein größeres Interesse für Beurteilung hat und nicht zuletzt auch durch ihre Teilnahme in der Gruppe gute Erfahrungen mit Bewertungsdiskussionen hat. Die Ergebnisse müssen mit Vorsicht interpretiert werden und können nur bedingt generalisiert werden, da es sich nicht um empirische Daten aus der gesamten Lehrerpopulation handelt.

immer geringer. Ob eine Beurteilung durch zwei Bewertende, die gemeinsam ein Urteil abgeben (*sambedömning*, etwa ein paralleles Bewertungsverfahren), oder durch zwei Bewertende, die ihre Ergebnisse nach der Beurteilung vergleichen (*medbedömning*, etwa eine Zweitkorrektur), verläuft, schien in dieser Studie, zumindest rein statistisch, eine geringere Rolle zu spielen.

Eine Tendenz zur Milde bzw. Strenge ist vorhanden, wenn Bewertende etwas strenger oder milder bewerten, unabhängig von der Qualität der faktischen Leistungen. Eine Tendenz zur Mitte kommt hingegen vor, wenn ein Bewertender Leistungen in der Mitte einstuft und dabei Schwierigkeiten hat, die jeweiligen zu bewertenden Leistungen voneinander zu trennen. In einer Studie von Eckes (2005) wurden Milde-Strenge-Tendenzen der Bewertenden bei einer Beurteilung fremdsprachlicher Leistungen untersucht. Der Test, eine standardisierte Version des *TestDaF* (*Test Deutsch als Fremdsprache*) wurde von 29 erfahrenen Bewertenden evaluiert. Das Ergebnis der Analyse zeigte, dass die Bewertenden sich bezüglich Strenge bzw. Milde untereinander deutlich unterschieden, aber dennoch eine akzeptable interne Konsistenz in ihren Gesamtbeurteilungen aufwiesen. Die Bewertenden wiesen zudem im Vergleich eine höhere interne Konsistenz im Hinblick auf die Gesamtbewertung der Leistungen auf als zu Konsistenzberechnungen einzelner Kriterien.

Inwiefern die Bewertererfahrung für die Tendenz zur Strenge bzw. Milde eine Bedeutung hat, ist in der Forschung untersucht worden. Die Studien haben jedoch keine eindeutigen Ergebnisse gezeigt. Während einige Befunde auf eine Tendenz zur Strenge unter Novizen hindeuten (vgl. Song & Caruso 1996), gibt es Studien, die das umgekehrte Verhältnis aufzeigen (vgl. Sweedler-Brown 1985; Barkaoui 2010a). Eine Erklärung dieser Diskrepanzen ergibt sich aus der Tatsache, dass unterschiedliche Bildungskontexte, Testteilnehmende und Bewerterkriterien vorlagen. Auch ein holistisches Bewertungsverfahren scheint hierbei eine Rolle zu spielen: Barkaoui (2010a) konnte in seiner Studie, in der 31 erfahrene Bewertende bzw. 29 Novizen Texte von Lernenden des Englischen als Zweitsprache beurteilten, eine Variabilität der beiden Bewertergruppen gerade bei der holistischen Beurteilung wahrnehmen. Die erfahrenen Bewertenden hatten bei der Bewertung im Vergleich zu den Novizen eine Tendenz zur Strenge. Sie neigten zudem dazu, sprachliche Korrektheit zu fokussieren und in höherem Ausmaß negative Kommentare zu geben. Demgegenüber gewichteten die Novizen in ihren Begründungen des Urteils stärker den Inhalt und verhielten sich positiver oder neutraler in den Kommentaren der holistischen Bewertung. Die Tatsache, dass erfahrene Bewertende einen Fokus auf sprachliche Korrektheit legen, könnte mehrere Erklärungen haben, z. B. die Tatsache, dass die erfahrenen Bewertenden häufig auch eine langjährige Erfahrung als Sprachlehrkräfte

haben (vgl. Song & Caruso 1996; Rinnert & Kobayashi 2001; Barkaoui 2010a). Erfahrungen im Hinblick auf das Testen und Beurteilen scheinen jedoch Bewertenden ein Zutrauen zu geben, kritisch zu bewerten und dies scheint auch zu zuverlässigeren Bewertungen zu führen (vgl. Sweedler-Brown 1985). Generell konnte jedoch in der Studie von Barkaoui auch eine große Variation innerhalb der Bewertergruppen gefunden werden, was darauf hinweist, dass der Einfluss von Bewertererfahrung allein die Bewertervariabilität nicht erklären kann.

Auch inwieweit holistische bzw. analytische Bewertungsansätze oder eine Bewertung ohne vorgeschriebene Bewertungsinstrumente (vgl. Bøhn 2016) zu bevorzugen sind, um zu einer möglichst hohen Beurteilerübereinstimmung zu gelangen, ist im Feld umstritten (vgl. Harsch & Martin 2013): Einige Untersuchungen zeigen Ergebnisse auf, die ein analytisches Verfahren begünstigen (vgl. Jönsson & Balan 2018), wohingegen es andererseits mehrere Studien zugunsten einer holistischen Beurteilung gibt (vgl. Barkaoui 2007; Graham et al. 2011). Befunde haben dennoch indiziert, dass eine hohe Beurteilerübereinstimmung bei einer holistischen Bewertung nicht immer bedeuten muss, dass die Bewertenden die Kriterien in ähnlicher Weise auffassen. Sie können aus verschiedenen Gründen zum selben Ergebnis gekommen sein (vgl. Lumley 2002; Harsch & Martin 2013).

Graham, Harris und Herbert (2011) haben eine Metastudie mit Fokus auf Effekte von formativen Bewertungen durchgeführt, die zugleich die Beurteilerübereinstimmung im Hinblick auf ein holistisches bzw. analytisches Verfahren bei der Bewertung von Textproduktionen aufklärt. In der Studie werden sowohl Konsensansätze (in etwa eine exakte Übereinstimmung der Bewertenden) als auch Konsistenzansätze (eine Korrelation zwischen den Bewertenden) dargestellt, wenn auch das jeweilige Maß für die Reliabilität nicht explizit angegeben wird. Sowohl die Konsistenzwerte als auch die Konsenswerte (sofern sie ermittelt werden) indizieren einen Vorrang für ein holistisches Bewertungsverfahren (ibid.).

Auch Barkaoui (2011a) kommt in einer Vergleichsstudie der beiden Bewertungsverfahren bei der Bewertung von Lernproduktionen auf Universitätsniveau zu dem Ergebnis, dass ein holistisches Bewertungsverfahren eine höhere Beurteiler-Reliabilität erreicht. Ein analytisches Verfahren hat wiederum eine höhere Intra-Rater-Reliabilität erreicht. Die Bewertenden hatten bei einem analytischen Verfahren zudem eine Tendenz, milder zu bewerten, möglicherweise, weil die Aspekte aufgeteilt waren. Eine wenig gute Leistung im Hinblick auf z. B. die grammatische Korrektheit deckt somit nur einen Teilaspekt und erhält nicht so einen starken Einfluss auf das Gesamtergebnis, wie es bei einer holistischen Bewertung der Fall sein könnte.

Darüber hinaus fanden sich Unterschiede bezüglich der Beurteilerübereinstimmung und der Sicherheit, mit welcher die Bewertenden fühlten, dass sie eine angemessene adäquate Note erteilten, die auf das Leistungsniveau der Lernerproduktionen zurückzuführen war. In einer Studie von Papageorgiou (2010) konnten Tendenzen unter Bewertenden wahrgenommen werden, dass sie oft nach eigenen Angaben Schwierigkeiten hatten, grenzwertige Leistungen, d. h. Leistungen, die die Anforderungen sehr knapp erfüllten, im Rahmen eines Standardsetting-Prozesses zu erfassen. Ähnliches haben Lehrkräfte bei der Beurteilung im Fach Deutsch als Fremdsprache in einem schwedischen Kontext angegeben: Sie finden es oft problematisch, die Grenze zwischen einer ausreichenden Note bzw. einer ungenügenden Note zu ziehen und beklagen, dass Bildungsdokumente und Prüfungsmaterialien eine ungenügende Unterstützung hinsichtlich dieser Problematik geben (vgl. Håkansson Ramberg 2021). Andererseits verweisen aber Quellen in der Forschung darauf, dass Bewertende es generell problematischer finden, Leistungen, die sich im mittleren oder höheren Bereich befinden, zu beurteilen (vgl. Birkel & Birkel 2002; Kuiken & Vedder 2014). Diese Befunde stehen im Einklang mit weiteren Studien hinsichtlich der Bewertung fremdsprachlicher Kompetenz aus einem schwedischen Schulkontext, in denen ebenfalls Schwierigkeiten bei der Bewertung von schriftlicher Sprachfertigkeit mittleren Noten wahrgenommen worden sind (vgl. Granfeldt & Ågren 2014).

Um zu validen Interpretationen von Lernerproduktionen zu gelangen, wird häufig die Bedeutung einer adäquaten Ausbildung im Hinblick auf das Testen und Bewerten hervorgehoben: „It is important, too, that human scorers are well-trained, so that they give similar scores to similar performances" (vgl. Douglas 2010: 27–28). Ein Faktor, der die Beurteilerkonsistenz beeinflussen kann, ist demzufolge Bewertertraining, was allerdings unter Novizen einen größeren Einfluss zu haben scheint (vgl. Weigle 1994). Anhand empirischer Daten wurde nachgewiesen, dass Bewertertrainings häufig positive Effekte haben (vgl. Weigle 1994; 1998; Berge 2005; Davis 2016), wobei die Qualität und Ausmaß des Bewertertrainings sowie kontextuelle Faktoren verständlicherweise die Ergebnisse beeinflussen. Dennoch schienen auch nach dem Bewertertraining gewisse Unterschiede zwischen Bewertenden weiterhin bestehen zu bleiben (vgl. Eckes 2008; Tengberg et al. 2017).

4.3 Sprachleistungsstudien mit Bezug auf die Referenzniveaus des GER

In Schweden gab es lange Zeit, mit Ausnahme der Bezugsstudie der Europäischen Kommission, ESLC, kaum empirische Studien über die Sprachkenntnisse

von Schülerinnen und Schülern im Hinblick auf die zweite Fremdsprache, die Bezug auf den GER nahmen. Darüber hinaus sind im Gegensatz zu Ländern wie z. B. Deutschland (DESI), Österreich (BIFIE) oder der Schweiz (Projekt HarmoS) außer den textuellen Vergleichsstudien von Skolverket (vgl. Kap. 2.4), keine landesweiten Erhebungen in diesem Bereich initiiert worden, die die Orientierung der Fremdsprachenstufen am GER untersuchen. Die wenigen vorhandenen Bezugsstudien zu diesem Thema hinsichtlich der zweiten Fremdsprache sind außerdem in Schweden vergleichsweise spät durchgeführt worden.

Im Hinblick auf die Beziehung zwischen den Fremdsprachenstufen des schwedischen Systems für die zweite Fremdsprache und den entsprechenden GER-Niveaus sind vor allem drei Untersuchungen in Schweden zu erwähnen, die im Folgenden genauer betrachtet werden: der Bezug des fakultativen Prüfungsmaterials hinsichtlich der ersten und zweiten Fremdsprache zum GER von Erickson (2011b; 2019), die Zuordnung mündlicher Kompetenzen in den drei Schulsprachen zum GER innerhalb des TAL-Projektes (vgl. Granfeldt et al. 2019b) sowie die Bezugsstudie von Aronsson (2020), die Lernprofile produktiver Fertigkeiten im Fach Spanisch untersucht. Die erste hier erwähnte Studie von Erickson untersucht folglich die Anbindung standardisierter Tests, wohingegen die beiden letzteren die Fremdsprachenkenntnisse von Lernenden am Ende der Grundschule fokussieren. An diesem Punkt soll die vorliegende Arbeit einen Beitrag leisten, die Beziehung von Lernerleistungen in Deutsch am Gymnasium zu den Referenzniveaus des GER besser zu verstehen.

Zu den zentralen Arbeiten unter den Bezugsstudien zählt insbesondere die Studie *European Survey on Language Competences* (ESLC) der Europäischen Kommission (vgl. European Commission 2012b). Diese Studie ist durchgeführt worden, um fremdsprachliche Kompetenzen in einer Auswahl von europäischen Ländern vergleichen zu können. Ein Ziel der Studie war es, das sprachliche Niveau europäischer Jugendlicher in den zwei meistgelernten Fremdsprachen in 16 teilnehmenden Bildungssystemen in Europa zu untersuchen, d. h. Englisch für alle Länder außer Großbritannien und in jedem Land die nach Englisch meistgewählte Fremdsprache. Allerdings ist zu bemerken, dass die mündliche Kompetenz der Lernenden in der ESLC-Studie nicht geprüft wurde. Die Studie zeigt, dass Lernende generell nicht die angestrebten GER-Niveaus erreichen, weder in der ersten noch der zweiten Fremdsprache, und dass viele Schülerinnen und Schüler darüber hinaus sogar nicht einmal die Anforderungen eines A1-Niveaus des GER erfüllen. Die Ergebnisse der Studie weisen jedoch auch auf sehr divergierende Ergebnisse hinsichtlich Sprachkompetenzen in den jeweiligen europäischen Ländern hin. In einigen Ländern, wie Schweden, konnten auch große Unterschiede zwischen der ersten bzw. zweiten

Fremdsprache (in Schweden: Englisch bzw. Spanisch) nachgewiesen werden (ibid.).

Die Ergebnisse dieser länderübergreifenden Lernstandserhebung der Europäischen Kommission haben in Schweden, wie bereits erwähnt, besondere Aufmerksamkeit erhalten. In der ESCL-Studie wurde in Schweden Sprachkompetenzen von Lernenden in Englisch und in der meistgewählten zweiten Fremdsprache Spanisch am Ende der Grundschule untersucht. Die Ergebnisse der Studie zeigen u. a., dass schwedische Schülerinnen und Schüler generell auf einem sehr hohen Niveau Englisch beherrschen. Bei der zweiten Fremdsprache Spanisch hingegen wiesen die schwedischen Jugendlichen schwächere Ergebnisse im Vergleich zu denen in Englisch auf. Zudem weisen die Ergebnisse im Vergleich zu französischen Lernenden in Spanisch, die jedoch aufgrund der geographischen und sprachtypologischen Nähe wahrscheinlich Vorteile haben, auf niedrigere Werte hin. In der Studie hat sich zudem herausgestellt, dass die große Mehrheit der schwedischen Spanischlernenden nicht das zu erwartende GER-Niveau (A2.1) erreichte.

Darüber hinaus haben viele Lernende auch das erste Niveau für die elementare Sprachverwendung (das A1-Niveau) nicht erreicht: 24 % der 15-Jährigen wurden beim Lesen, 37 % beim Hören und ganze 45 %, also fast die Hälfte der schwedischen Schülerinnen und Schüler in Spanisch, beim Schreiben auf ein Niveau unterhalb von A1 (sog. Pre-A1-Niveau) eingeordnet (ibid.). Da anzunehmen ist, dass sich Kompetenzen unterschiedlich schnell entwickeln und dass rezeptive Fertigkeiten häufig ein höheres Niveau im Vergleich zu den produktiven zeigen (vgl. Tschirner 2008), scheint das auch im Hinblick auf die Fremdsprachenkenntnisse für viele der Schülerinnen und Schüler in dieser Studie zuzutreffen. Gemäß den Ergebnissen der Studie befinden sich in etwa 90 % der schwedischen 15-Jährigen beim Schreiben in Spanisch unter dem angestrebten A2.1-Niveau des GER. Laut der Studie erreichten nur etwa 10 % der Schülerinnen und Schüler am Ende der Grundschule mehr als ein A1-Niveau in Spanisch bei einem Test des schriftlichen Ausdrucks (European Commission 2012b: 235). Die Ergebnisse implizieren somit insgesamt, dass eine sehr große Anzahl der schwedischen 15-Jähringen am Ende der Grundschule die Anforderungen des zu erwartende A2-Niveaus in Spanisch nicht erfüllen, insbesondere scheint dies für die schriftliche Kompetenz der Fall zu sein.

Mögliche Erklärungen für die niedrigeren Ergebnisse für Spanisch im Vergleich zu Englisch sind u. a., dass die Lernenden außerhalb des Klassenzimmers selten mit Spanisch in Kontakt kommen und dass Lehrkräfte in Spanisch zu dieser Zeit weniger oft eine pädagogische Ausbildung hatten als Lehrkräfte in Englisch. In einer Studie von Riis und Francia (2013) konnte gezeigt werden,

dass ein großer Anteil der Lehrkräfte in Spanisch weder die für das Niveau angeforderte pädagogische Ausbildung absolviert hatten noch generell die Anforderungen für einen Nachweis eines abgeschlossenen Lehramtsstudiums (d. h. ihre sog. Legitimation für Lehrkräfte, *lärarlegitimation*) erfüllten.[68] Inwiefern ähnliche Zustände ebenso für die mündliche Kompetenz in Spanisch und für die Sprachkompetenzen der 15-Jährigen in den Fremdsprachen Deutsch und Französisch im schwedischen Bildungssystem vorliegen, kann allerdings auf Basis dieser Untersuchung leider nicht geklärt werden.

Auch wenn nur wenige empirische Studien zur Fremdsprachenkompetenz von Schülerinnen und Schülern im schwedischen Schulkontext durchgeführt worden sind, wurde festgelegt, dass die standardisierten Tests in etwa den zu erwartenden Niveaus entsprechen. In zwei Studien wurde von Erickson (2011b; 2019) die Anbindung von Prüfungen des schriftlichen Ausdrucks im Fach Englisch und in den zweiten Fremdsprachen Deutsch, Französisch und Spanisch an die Referenzniveaus des GER im Hinblick auf Kriterien und Inhalt untersucht. Hierbei wurden internationale GER-Experten gebeten, das nationale Testmaterial in Bezug auf die Niveaus im Referenzrahmen einzuordnen. Die Ergebnisse dieser Untersuchungen haben insgesamt das Resultat der vorherigen textuellen Analysen (vgl. Kap. 2.4.2) bestätigt und darauf hingewiesen, dass die nationalen Testmaterialien am Ende der Grundschule für Englisch generell einem erreichten Niveau B1.1 und für die zweite Fremdsprache einem erreichten A2.1-Niveau entsprechen. Zu erwähnen ist allerdings, dass Beispiele mündlicher und schriftlicher Leistungen mit niedrigeren Benotungen im Material nicht immer auf das zu erwartende Niveau, sondern ein Niveau niedriger eingestuft wurde, während Leistungsbeispiele mit höherer Benotung oft auf höheren GER-Stufen eingeordnet wurden (Erickson 2019). Inwiefern dies auch die realen Fremdsprachenkompetenzen der Schülerinnen und Schüler widerspiegelt wurde nicht untersucht. Die Tatsache, dass die Prüfungen dem zu erwartenden Niveau entsprechen ist vom Qualitätsstandpunkt her betrachtet sehr gut, sollte jedoch durch empirische Analysen der Sprachkompetenzen der schwedischen Schülerinnen und Schüler unter realen Verhältnissen ergänzt werden.

Eine schwedische Studie versucht die Forschungslücke der ESLC-Studie hinsichtlich der mündlichen Kompetenz (vgl. European Commission 2012b) zu

68 Der Anteil von Spanisch-Lehrkräften mit pädagogischer Ausbildung und der damit verbundenen Legitimation hat sich in Schweden zwar seitdem wesentlich erhöht, liegt aber immer noch für Lehrkräfte in den Fächern Deutsch und Französisch, sowohl in der Grundschule als auch am Gymnasium, deutlicher höher (vgl. Skolverket 2019b).

schließen. An schwedischen Grundschulen wurden Daten mündlicher Sprachfertigkeit von Schülerinnen und Schülern in einer groß angelegten Studie erhoben und dabei wurden alle drei größeren Fremdsprachen des Bildungssystems in Schweden involviert. Diese Untersuchung ist ein Teil des *TAL-Projekts* (*TAL-Project: Teaching, Assessment and Learning of second foreign languages*), einer größeren Forschungsstudie, die den Sprachunterricht und die Bewertung mündlicher Sprachkompetenz im Fach *Moderna språk* bei Lernenden der neunten Jahrgangsstufe evaluiert. Im Rahmen der Studie wurde u. a. untersucht, in welchem Ausmaß Schülerinnen und Schüler in ihren zweiten Fremdsprachen Deutsch, Französisch oder Spanisch das erwartete Referenzniveau A2.1 für die mündliche Sprachfertigkeit erreichten. Das Ergebnis der Studie zeigt, dass weniger als die Hälfte der schwedischen 15-Jährigen das A2.1-Niveau erreichen. Für die Fremdsprache Deutsch ist das Niveau der mündlichen Kompetenz der Lernenden allerdings höher als in den anderen getesteten Sprachen. Es könnte jedoch mehrere Erklärungen für die niedrigeren Ergebnisse geben, u. a., dass die Schülerinnen und Schüler die Bedingungen bei der Prüfung, z. B. ohne jegliche schriftliche Hilfsmittel zu sprechen, nicht gewohnt waren und sich in der Anwesenheit der Wissenschaftler verunsichert fühlten (vgl. Granfeldt et al. 2019b). Inwiefern die niedrigeren Ergebnisse der schwedischen Schülerinnen und Schüler am Ende der neunten Jahrgangsstufe bezüglich der mündlichen Kompetenz auch für die schriftliche Kompetenz in Deutsch gelten, wurde im Rahmen dieser Studie nicht untersucht.

Auch die Ergebnisse einer der wenigen empirischen Studien deuten darauf hin, dass Bewertungen mündlicher und schriftlicher Leistungen von schwedischer Lernenden in Spanisch kein zufriedenstellendes Ergebnis geben. In dieser Studie von Aronsson (2020) wurde der Bezug mündlicher bzw. schriftlicher Kompetenz kurz nach Ende der Grundschule für die Fremdsprachenstufe *Spanska 2* („Spanisch 2") zu bestimmten GER-Niveaus untersucht. Die insgesamt 90 Lernerproduktionen wurden von jeweils zwei schwedischen Bewertenden und zwei GER-Bewertenden beurteilt. Aronsson gelangt in ihrer Studie zu folgendem Schluss: Ein sehr großer Anteil der Lernenden erreicht am Ende der Grundschule nicht das intendierten GER-Niveau A2.1 für Schreiben und Sprechen in Spanisch. Des Weiteren haben lediglich die Schülerleistungen mit höheren Noten (d. h. die Noten C–A) das GER-Niveau A2.1 erreicht. Keine der Lernproduktionen mit einer ausreichenden Note E, die in etwa dem Mindestniveau eines erreichten GER-Niveaus A2.1 entsprechen sollten, hat demzufolge das angestrebte Niveau erreicht. Die Studie zeigt zudem, dass die Lernenden generell eine höhere Schreibkompetenz in Spanisch im Vergleich zu der mündlichen Kompetenz aufweisen. Insgesamt befanden sich laut der Studie circa

11 % der mündlichen Lernproduktionen und 26 % der Textproduktionen auf einem A2.1-Niveau, ein Ergebnis, das im Vergleich zur ESLC-Studie für die schriftliche Kompetenz zwar auf eine Verbesserung hinweist, jedoch weit entfernt von dem erwarteten Referenzniveau des GER ist.

Zu beachten ist aber, dass die Lernenden in dieser Studie von Aronsson sich bei der Datenerhebung im ersten Jahrgang am Gymnasium und nicht am Ende der neunten Jahrgangsstufe in der Grundschule befanden. Es kann angenommen werden, dass Lernende, die in der Grundschule eine nicht ausreichende Note F erhalten haben, am Gymnasium mit ihrer in der Grundschule gewählten Sprache nicht fortgefahren sind und eine andere Sprache statt Spanisch gewählt haben. Diese Lernenden sind möglicherweise daher nicht in der Untersuchung dabei. Aus diesem Grund kann vermutet werden, dass der Anteil von Lernenden inklusive dieser Gruppe, der am Ende der neunten Jahrgangsstufe die Anforderungen eines A2.1-Niveau in Spanisch erreicht hätte, noch niedriger geworden wäre.

Allen Studien ist gemeinsam, dass Bewertungen von Sprachkompetenzen am Ende der neunten Jahrgangsstufe oder für dieses Niveau bestimmte Testmaterialien im Fokus standen. Deutlich weniger empirische Bezugsstudien zum GER liegen bislang für die Fremdsprachenstufen am Gymnasium vor. Hier können aber zwei Studien mit Fokus auf Lernergebnisse in Englisch herangezogen werden. In einer Validierungsstudie aus dem Jahr 2002 hat Tyllered in einem internen Bericht der schwedischen Schulbehörde eine sehr gute Übereinstimmung zwischen dem Prüfungsmaterial der Fremdsprachenstufe *Engelska 7* („Englisch 7“) für die produktiven Fertigkeiten Sprechen und Schreiben und dem *Cambridge Certificate in Advanced English* (CAE), in etwa einem GER-Niveau C1, nachweisen können (Tyllered 2002).[69] Des Weiteren hat Borger (2018) eine ungefähre Relation zwischen dem Mindestniveau für die Fremdsprachenstufe *Engelska 6* („Englisch 6“) des schwedischen Systems und einem GER-Niveau B2.1 hinsichtlich der mündlichen Kompetenz festgestellt. Diese Übereinstimmungen stimmen generell mit den angestrebten GER-Niveaus für diese Fremdsprachenstufen in Englisch überein. Die Tatsache, dass schwedische Lernenden am Gymnasium die intendierten Niveaus in Englisch erreichen

69 Auch die Kurse in der Grundschule und am Gymnasium in Englisch gehören zum gemeinsamen System für Fremdsprachen in Schweden. Am Gymnasium können drei Kurse in Englisch belegt werden, *Engelska 5*, *Engelska 6* und *Engelska 7*, deren jeweiligen Mindestniveaus in etwa die GER-Niveaus B1.2, B2.1 und B2.2 entsprechen (vgl. Kap. 2.4.2, Tab. 6).

ist jedoch kaum überraschend, da bereits die internationale ESLC-Sprachstudie gezeigt hatte, dass sich schwedische Schülerinnen und Schüler bereits in der Grundschule durch eine hohe Kompetenz in Englisch auszeichnen. Bisher gibt es in Schweden dahingegen noch keine Bezugsstudie zum GER, die Sprachkompetenzen für die zweite Fremdsprache am Gymnasium in Betracht gezogen hat.

Neben den erwähnten Studien aus dem schwedischen Kontext gibt es eine wachsende Zahl von Untersuchungen, die sich explizit auf die Referenzniveaus des GER stützen, um Fremdsprachenkenntnisse von Schülerinnen und Schülern innerhalb von Bildungssystemen unterschiedlicher Länder auszuwerten und zu beschreiben. Zu nennen ist eine Untersuchung von Sprachkompetenzen in den Schulfächern Deutsch und Englisch am Ende der neunten Jahrgangsstufe, jedoch ausschließlich in einem deutschen Schulkontext, die große deutsche Schulleistungsstudie, *Deutsch-Englisch-Schülerleistungen-International* (DESI). Das DESI-Konsortium kommt zu dem Ergebnis, dass die Mehrheit der Schülerleistungen mindestens das für den Hauptschulabschluss erwartete A2-Niveau nach der neunten Jahrgangsstufe in Englisch erreicht hat, was im Einklang mit den deutschen Bildungsstandards ist, wobei auch in gewissem Ausmaß Anforderungen für höhere GER-Niveaus erfüllt wurden (vgl. DESI-Konsortium 2006).[70] Da die Voraussetzungen für das Erlernen von Englisch aber anders sind als bei einer zweiten Fremdsprache und zudem die Situation in einem deutschen Kontext untersucht wurde, hat diese Studie für die vorliegende Untersuchung weniger Relevanz.

Zu nennen sind in diesem Zusammenhang u. a. auch eine Studie der fremdsprachlichen Kompetenz des Englischen und Schwedischen bei 15-Jährigen aus Finnland (vgl. Hildén et al. 2019), eine Untersuchung der L2-Kompetenz (Deutsch/Italienisch) Südtiroler Schülerinnen und Schüler im Alter zwischen 17 und 18 Jahren (im KOLIPSI-Projekt: vgl. Abel et al. 2012) sowie eine Auswertung der Sprachkompetenz in Englisch von Schülerinnen und Schülern der 8. Klassen in Österreich (vgl. BIFIE 2012). Ebenso zu erwähnen ist das schweizerische Projekt HarmoS (vgl. Lenz & Studer 2008; Schneider et al. 2009), welches

70 Die in der Studie untersuchten mündlichen Sprachproduktionen in Englisch konnten durch einen kommerziellen Test laut den Wissenschaftlern auf den Europäischen Referenzrahmen und die Bildungsstandards bezogen werden (vgl. Klieme 2006). Eine Einschränkung in diesem Zusammenhang ist allerdings, wie De Florio Hansen (2015: 42) bemerkt, dass der in der DESI-Studie verwendete kommerzielle Test nicht explizit auf die Referenzniveaus des GER bezogen war, und die Ergebnisse sollten daher mit Vorsicht interpretiert werden.

zur Harmonisierung und Entwicklung der nationalen Bildungsstandards in den Fremdsprachenfächern basierend auf empirischen Untersuchungen den GER als Referenzpunkt verwendet hat. Diese Quellen bieten für die vorliegende Arbeit jedoch ebenfalls Anhaltspunkte für den GER-Bezug, sie stellen aber verschiedene Bildungskontexte dar, u. a. im Hinblick auf Jahrgangsstufen und Unterrichtsstunden.

4.4 Fazit

In diesem Kapitel wurden Studien im Hinblick auf beurteilergestützte Bewertung aus verschiedenen Perspektiven vorgestellt. Diese Studien dienen somit als Grundlage der vorliegenden Studie. Eine wesentliche Begrenzung hierbei ist jedoch, dass bei vielen der bisherigen Studien Englisch als Zweit- oder Fremdsprache oder die Muttersprache (L1) im Fokus standen und Untersuchungen anderer Fremdsprachen außer Englisch trotz zunehmender Aufmerksamkeit für L3-Forschung in den letzten Jahren (vgl. Bardel et al. 2016) kaum zu finden sind. Wenn mehrere Studien, trotz unterschiedlicher Kontextbedingungen, ähnliche Befunde zeigten, kann aus diesen Ergebnissen dennoch eine greifbare Generalisierbarkeit angenommen werden und sie können somit für einen schwedischen Schulkontext Relevanz haben.

Viele Studien aus einem schwedischen Kontext zeigen, dass beurteilergestützte Bewertung ein komplexer Prozess ist, wobei Unterschiede im Hinblick auf die Interpretation und das Verstehen von Kriterien der Bewertenden zu finden sind. Daraus ergibt sich, dass gewisse Aspekte in Bewerterurteilen von Bewertenden oft mehr Gewicht erhalten. Weitgehend überwiegen in den Bewerterurteilen mehrerer Studien häufig Aspekte der linguistischen Kompetenz. Generell scheinen unterschiedliche Faktoren, wie Hintergrundvariabeln der Bewertenden wie *Ausbildungshintergrund*, *Grad an Unterrichts- und Bewertererfahrungen*, *Muttersprache* und *Alter* sowie unterschiedliche *kontextuell* bedingte Bewerterkulturen, das *Bewerterverfahren* sowie *Merkmale der Leistungen*, einen Einfluss bei der Beurteilung ausüben zu können.

Zudem wurde wahrgenommen, dass Bewertende Unterschiede bezüglich Strenge-Milde-Tendenzen und Bewerterprofilen aufweisen. Viele dieser vorherigen Studien zur Bewertervariabilität konnten zeigen, dass Bewertererfahrung bzw. vorangegangene Bewertertrainings einen positiven Effekt haben. Des Weiteren können verschiedene Herangehensweisen bei der Beurteilung, wie holistische bzw. analytische Bewertungsverfahren, zu Unterschieden in den Bewerterurteilen führen. Aus diesem Grund sind auch die unterschiedlichen Bewertungsverfahren der Lehrkräfte von Interesse. Da zudem eine hohe

Übereinstimmung bei einer holistischen Beurteilung eine Nichtübereinstimmung bezüglich der zu bewertenden Aspekte verbergen kann, ist von Gewicht, dass auch diese Perspektive bei der Beurteilung untersucht wird. Auch wenn der Fokus der Berichte im schwedischen Kontext häufig auf der Bewerterübereinstimmung liegt, gibt es nur wenige Studien zur Bewerterkonsistenz bezüglich der Bewertung in einer Fremdsprache aus einer schwedischen Perspektive.

Insgesamt zeigen bisherige Studien, dass Aufsätze im Fach Schwedisch (die Mehrheitssprache) eine niedrigere Konsistenz im Vergleich zu Konsistenzwerten für die Beurteilung von Texten im Fach Englisch (die erste Fremdsprache) aufweisen. Darüber hinaus zeigen Studien zur Beurteilung durch die eigene Lehrkraft, auch wenn die Ergebnisse nicht immer eindeutig sind, gewisse Milde-Tendenzen bei der Bewertung auf. Aufmerksamkeit im schulischen Kontext haben die Kontrollkorrekturen der schwedischen Schulaufsichtsbehörde erhalten, da die Ergebnisse auf eine zum Teil große Variabilität der Beurteilungen schwedischer Lehrkräfte hinweisen, z. B. dass die eigene Lehrkraft eine etwas höhere Benotung gibt (vgl. Skolinspektionen 2018). Allerdings sollten die fragwürdigen Methoden beim unabhängigen Benotungsverfahren (vgl. Gustafsson & Erickson 2013) in Frage gestellt werden. Die Bewertung fremdsprachlicher Kompetenz im Fach Deutsch ist jedoch bisher nicht systematisch untersucht worden.

Studien zur Einordnung von Sprachkompetenzen im Fremdsprachenbereich nehmen in immer höherem Grad Bezug auf den Gemeinsamen europäischen Referenzrahmen. Hierbei sind viele Studien bei der Einstufung von Testergebnissen internationaler und nationaler Prüfungen am GER ausgerichtet, wobei in den letzten Jahren ein erhöhtes Interesse für Kompetenzmessungen und Vergleiche von Bildungssystemen unterschiedlicher Länder verzeichnet werden kann, die sich an den Referenzniveaus des GER orientieren. Hierzu gehört auch die zunehmende Anzahl von Studien zur Zuordnung von Fremdsprachenkenntnissen von Schülerinnen und Schülern in den Bildungssystemen der jeweiligen Länder oder Regionen zu den Sprachniveaus des GER. Einige wenige Studien haben dennoch Fremdsprachenkenntnisse in der zweiten Fremdsprache von Schülerinnen und Schülern im schwedischen Bildungssystem in Bezug auf den Referenzrahmen empirisch untersucht. In diesen bisher durchgeführten empirischen Studien hat sich erwiesen, dass sich Sprachlernende der zweiten Fremdsprache in der schwedischen Grundschule nicht immer auf dem zu erwartenden sprachlichen Niveau befinden. Dies scheint vor allem im Hinblick auf die produktiven Fertigkeiten, d. h. die mündliche und schriftliche Kompetenz, der Fall zu sein. Der Bezug zum GER hinsichtlich der schriftlichen

Kompetenz im Fach Deutsch wurde bislang im schwedischen Schulkontext nicht untersucht.

Der Überblick zeigt, dass nur wenige wissenschaftliche Studien zur Bewertung einer zweiten Fremdsprache in einem schwedischen Schulkontext zu finden sind und dass ein Desiderat nach deutschdidaktischer Forschung im Hinblick auf die Bewertung von Schülerleistungen auf Grundlage empirischer Daten vorliegt. Es ist demnach auch wünschenswert, relevante Aspekte der Validität bei einer Bewertung fremdsprachlicher Kompetenz stärker in den Fokus zu nehmen.

5. Forschungsdesign und Forschungsmethodik

In diesem Kapitel werden das Forschungsdesign und grundlegende methodische Aspekte der vorliegenden Studie beschrieben und begründet. Zunächst erfolgt eine kurze Erläuterung zum Entwurf der Forschungsvorgehensweise und der Mixed-Methods-Ansätze, an welchen sich die Forschungsmethodik der vorliegenden Studie orientiert (Kap. 5.1). Danach folgen eine Beschreibung und Kontextualisierung im Hinblick auf die Methoden der Datenerhebung (Kap. 5.2). Anschließend wird auf die Methodik der qualitativen bzw. quantitativen Datenanalyse eingegangen (Kap. 5.3). Abschließend werden Begrenzungen bezüglich der Forschungsmethodik erörtert (Kap. 5.4).

5.1 Orientierung an Mixed-Methods-Ansätzen

Die vorliegende Untersuchung gliedert sich in unterschiedliche Teiluntersuchungen und besteht aus mehreren Phasen. Um Aspekte der Validität im Hinblick auf die Bewertung schriftlicher Kompetenz untersuchen zu können, werden hierbei sowohl qualitative als auch quantitative Methoden herangezogen. Wie bereits in den vorherigen Kapiteln gezeigt wurde, werden im Bereich des Fremdsprachentestens vorwiegend quantitative Methoden zum Einsatz gebracht, während qualitativ orientierte Studien weniger oft vorkommen. Eine Orientierung an sog. Mixed-Methods-Ansätzen ermöglicht indessen die Erhebung qualitativer und quantitativer Daten. Ein Mixed-Methods-Verfahren stellt jedoch keine instrumentalisierten Richtlinien dar, vielmehr handelt es sich um einen Leitfaden für das Forschungsdesign und die Interpretation der Ergebnisse. Durch ein Zusammenführen qualitativer und quantitativer Datenerhebung und deren Analyse in derselben Studie können die Stärken beider Methoden ein tieferes Verständnis für das studierte Phänomen im Vergleich zu lediglich einer Forschungsmethode bieten.

Mittlerweile werden Mixed-Methods-Ansätze, vor allem von amerikanischen Forschern, als ein drittes methodologisches Paradigma betrachtet (vgl. Johnson & Onwuegbuzie, 2004; Kuckartz 2014a). Studien mit einem Mixed-Methods-Design kommen in erster Linie in den Sozial- und Erziehungswissenschaften und vor allem im angelsächsischen Raum vor (Kuckartz 2014a). Jedoch breitet sich ihre Anwendung auch darüber hinaus immer weiter in Richtung einer global verwendeten Methode aus. Mit der gegenwärtigen Expansion dieser Methode wird das Mixed-Methods-Verfahren zunehmend

ausgearbeitet und an das jeweilige Forschungsgebiet angepasst (ibid.). Bei der Wahl zwischen unterschiedlichen Mixed-Methods-Designformen müssen gewisse Überlegungen vorgenommen werden, z. B. was man durch die Kombination zweier Methoden gewinnt, in welcher Reihenfolge die qualitative bzw. quantitative Datenerhebung durchgeführt wird und zu welchem Zeitpunkt im Forschungsverlauf die qualitativen bzw. quantitativen Daten integriert werden (vgl. Kuckartz 2014a: 57–76).

Da in diesem Falle sowohl qualitative als auch quantitative Methoden zur Beantwortung der Forschungsfragen geeignet sind, orientiert sich die vorliegende explorative Arbeit an Mixed-Methods-Ansätzen hinsichtlich Forschungsdesign und Forschungsmethodik. Die Studie basiert dabei auf einem konvergenten parallelen Design, eines der am häufigsten verwendeten Designs innerhalb von Mixed-Methods-Studien im Bereich des Fremdsprachentestens (vgl. Jang et al. 2014; für einen Überblick über Mixed-Methods-Designs vgl. Ziegler & Kang 2016). Bei einem Parallel-Design laufen eine qualitative und eine quantitative Teilstudie parallel und unabhängig voneinander ab. Das Ziel einer konvergenten parallelen Designstudie ist es, die Ergebnisse der qualitativen und quantitativen Auswertungsmethoden miteinander zu vergleichen und in Relation zu setzen.

Die praktische Realisierung des vorliegenden Forschungsprojektes kann in folgende vier Phasen gegliedert werden, siehe Abb. 7:

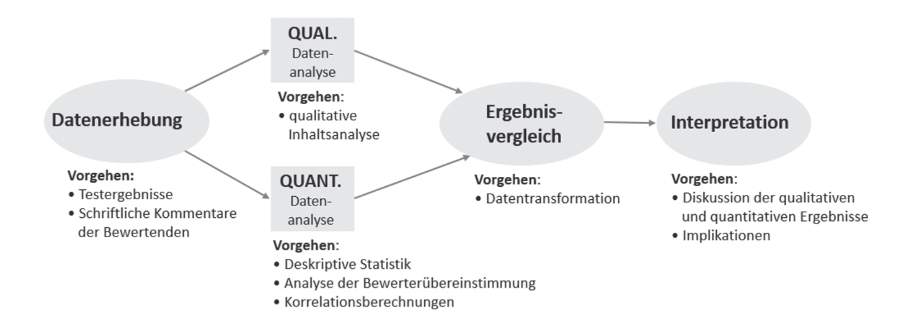


Abb. 7: Ablaufschema des parallelen Forschungsdesigns

In der ersten Phase bildet das Projekt bei der Planung und bei der *Datenerhebung* eine Einheit, wobei Textproduktionen schwedischer Schülerinnen und Schüler von unterschiedlichen Bewertenden evaluiert werden und das Material der Studie in Form von Testergebissen und schriftlichen Bewerterkommentaren gesammelt wird. Danach trennen sich die Wege, und die *qualitative* bzw.

die *quantitative Datenanalyse* verlaufen unabhängig voneinander. Hierfür wird einerseits eine qualitative Inhaltsanalyse durchgeführt und andererseits werden quantitative Berechnungen zur deskriptiven Statistik, zur Bestimmung der Bewerterübereinstimmung sowie Korrelationsanalysen vollzogen. Die Ergebnisse der qualitativen und die quantitativen Auswertungsmethoden werden in der dritten und vierten Phase des Forschungsverlaufs aufeinander bezogen (*Ergebnisvergleich*) und schließlich interpretiert (*Interpretation*). Dieses Forschungsdesign ist für die vorliegende Studie gewählt worden, um unterschiedliche Aspekte der Validität (vgl. Weir 2005) evaluieren zu können, dabei der Gliederung einer Validierung nach einem argumentbasierten Ansatz (z. B. Kane 2013) folgend. Für die vorliegende Studie bedeutet dies, dass nicht nur die Ergebnisse der Bewertungen (*das Produkt*) untersucht werden, sondern auch das Verständnis der jeweiligen Bewertenden über das zu messende Konstrukt und inwiefern sich dies in ihren Urteilen ähnlich oder divergierend wiederfindet (*der Prozess*). Das Parallel-Design lässt somit unterschiedliche Perspektiven ans Licht kommen und ermöglicht dadurch ein komplexeres Bild des vorliegenden Forschungsproblems. Die Vorgehensweise bei der empirischen Datenerhebung und den Analysen im Sinne des Mixed-Methods-Ansatzes wird im Folgenden näher beschrieben.

5.2 Datenerhebung

In diesem Kapitel werden die Teilnehmenden der Studie, der verwendete Test und die jeweiligen Bewertungsskalen näher beschrieben. Die schriftlichen Schülerproduktionen wurden über einen Test des schriftlichen Ausdrucks an mehreren Gymnasialschulen in Schweden erhoben und nach einem Auswahlverfahren in einem Korpus zusammengestellt. Diese Schülertexte wurden von sowohl schwedischen Bewertenden als auch GER-Bewertenden evaluiert. Die daraus abgeleiteten Daten in Form von Testergebnissen und Bewerterkommentaren bilden als Untersuchungsgrundlage die Primärdaten in der vorliegenden Arbeit.

Probandenpopulation

Um die Fragestellungen der vorliegenden Arbeit beantworten zu können, wurden Schülerleistungen an verschiedenen Schulen gesammelt. Die Wahl für die zu untersuchenden Stufen fiel auf die Fremdsprachenstufen *Tyska 3*, *Tyska 4* und *Tyska 5,* die gemäß dem Kommentarmaterial der schwedischen Lehrpläne für Fremdsprachen in etwa mit den Referenzniveaus A2.2, B1.1 und B1.2 des GER vergleichbar sind (vgl. Skolverket 2011b). Zahlreiche schwedische Schülerinnen

und Schüler besuchen diese Kurse im Gymnasium, vor allem *Tyska 3* und *Tyska 4*, und zudem kann angenommen werden, dass die Deutschlernenden dieser Stufen bereits ein ausreichendes Sprachniveau erreicht haben, auf dem ihre schriftlichen Leistungen mit einem vergleichbaren Referenzniveau des GER in Relation gesetzt werden können. Gemäß der schwedischen Schulbehörde entspricht das Sprachniveau von Deutschlernenden, die den Kurs auf der Stufe *Tyska 5* mit einer bestandenen Note abgeschlossen haben, einem erfüllten B1-Niveau (ibid.), weshalb dieses Niveau für den folgenden Vergleich ausgewählt wurde (vgl. auch Kap. 2.4.2).

Die Probanden waren schwedische Schülerinnen und Schüler am Ende der ersten, zweiten und dritten Jahrgangsstufe des schwedischen Gymnasiums.[71] Bei der Datenerhebung wurden die Empfehlungen zu forschungsethischen Grundsätzen des schwedischen Forschungsrats *Vetenskapsrådet* (2002) beachtet. Die wichtigsten Prinzipien sind hierbei die Informationspflicht, die Anforderung bewusster Zustimmung zum Forschungsprojekt sowie die Grundsätze zu Vertraulichkeit, Integrität und Nutzung von personenbezogenen Daten. Die Teilnehmenden hatten schriftliche Informationen über die Studie bekommen und mussten ihr Einverständnis dazu abgeben, dass ihre Texte für Forschungszwecke verwendet werden dürfen. Zudem wurden sämtliche Prüfungsteilnehmende über die Zielsetzung der Studie und darüber, dass sie jederzeit ihr Einverständnis zur Mitwirkung an der Studie zurückziehen konnten, informiert. Sie wurden auch darüber in Kenntnis gesetzt, dass ihre Schülerleistungen jeweils mit einem Kode gekennzeichnet würden und dass die Namen der Schülerinnen und Schüler, der Schulen und eventuelle Wohnorte ausgelassen würden, um die Anonymität der Probanden zu gewährleisten.

Die Informationen über die Studie waren auf Schwedisch verfasst, damit sie für die Probanden leicht zu verstehen waren. Des Weiteren enthielten sie die Kontaktangaben der Wissenschaftlerin. Da die teilnehmenden Schülerinnen und Schüler nicht unter 15 Jahre alt waren, war es den Prinzipien des schwedischen Forschungsrates folgend nicht notwendig, eine Erlaubnis von den Erziehungsberechtigten einzuholen. Mit wenigen Ausnahmen haben die potenziellen Probanden einer Teilnahme an der Studie zugestimmt. Jeder schriftlichen Schülerleistung wurde eine kombinierte Buchstaben-Zahlenkennung zugeteilt. Es ist daher nicht möglich, die Teilnehmenden der Studie zu identifizieren. Das Material sowie die Namen der Probanden wurden im Einklang

71 D. h. etwa aus der 11., 12. und 13. Klasse eines deutschen Gymnasiums.

mit den Regeln zur Datenschutz-Grundverordnung, GDPR, von der Wissenschaftlerin vertraulich behandelt.

Um zusätzliche Informationen über den Hintergrund der Prüfungsteilnehmenden zu erlangen, die für die spätere Analyse wichtig sein könnten, füllten sie im Anschluss an die Prüfung einen Fragenbogen auf Schwedisch aus. Die teilnehmenden Probanden begannen in der Grundschule, ab der sechsten oder siebten Klasse, mit Deutsch als Fremdsprache.[72] In 4 der 25 Schülergruppen hatten Deutschlernende bereits Sprachprüfungen, die sich am GER orientieren, abgelegt. In drei Gruppen handelte es sich um das *Deutsche Sprachdiplom der Kulturministerkonferenz* (DSD) und in einer Gruppe um das *Goethe-Zertifikat B1.*[73] Die Probanden belegten theoretische Studienausrichtungen am Gymnasium und dabei überwiegend Ausbildungsprogramme mit naturwissenschaftlicher oder geistes- und sozialwissenschaftlicher Ausrichtung. Aber auch Lernende, die Ausbildungsprogramme mit wirtschaftlicher oder ästhetischer Spezialisierung sowie einem Schwerpunkt auf Sprachen besuchten, sind im Material vertreten.

Prüfungsmaterial und Aufgabenstellung

Bei dem empirischen Testverfahren der vorliegenden Studie ist von großer Bedeutung, dass ein Test verwendet wird, der valide und zuverlässige Aussagen sowohl über die zu messende sprachliche Kompetenz als auch über den Bezug zum Europäischen Referenzrahmen zulässt. Zum einen muss sichergestellt werden, dass der Test dem fokussierten Sprachniveau des GER angemessen ist (vgl. Council of Europe 2009), Zum anderen ist weithin bekannt, dass die Aufgabenstellung eine Auswirkung bei der Bewertung haben kann (vgl. Weir 2005). Die schriftliche Sprachfertigkeit in einer Fremdsprache wird jedoch häufig durch standardisierte Prüfungen getestet, wobei die Lernenden danach ihre Sprachkenntnisse in Form eines Zertifikats oder Sprachdiploms nachweisen können.

---

72 In der vorliegenden Studie gab es dennoch auch Probanden, die in anderen Jahrgängen mit Deutsch angefangen hatten. Das beruhte darauf, dass sie früher in einem anderen Land zur Schule gegangen sind (z. B. in Norwegen) oder dass sie eine Schule mit einem deutschen Profil besucht hatten.

73 Die Tatsache, dass Prüfungsteilnehmende bereits eine Zertifikatsprüfung belegt haben, wird in diesem Zusammenhang nicht als Problem gesehen, da die Bewertungen und nicht die Probanden im Zentrum dieser Arbeit stehen. Allerdings könnte diese Kenntnis eventuell die Deutschlehrkräfte bei der Bewertung beeinflussen.

Bisherige Studien haben feststellen können, dass der GER sich als Referenzpunkt und Basis für Leistungsmessungen eignen kann (z. B. DESI-Konsortium 2006; European Commission 2012b) und dass anerkannte Zertifikatsprüfungen als Einstufungsinstrument verwendet werden können, um den Sprachstand von Lernenden festzustellen (vgl. Goertler et al. 2018). Es hat sich zudem gezeigt, dass erfahrene Bewertende sehr reliable Beurteilungen im Hinblick auf die GER-Stufen leisten können (vgl. Tschirner & Bärenfänger 2012).

Bei Tests dieser Art können zwei Ansätze unterschieden werden: die Leistungen werden anhand einer Skala beurteilt, die entweder mehrere Stufen oder Niveaus der Sprachkompetenz umfasst oder die auf ein spezifisches Niveau ausgerichtet ist. In Bezug auf den ersterwähnten Fall wird von einem *Multi-Level-Ansatz* gesprochen, während der letztgenannte Fall unter der Bezeichnung *Uni-level-Ansatz* oder niveauspezifischer Ansatz firmiert (vgl. Harsch & Rupp 2011; Grotjahn 2017). Harsch und Rupp (2011) bevorzugen einen niveauspezifischen Ansatz, wenn evaluiert werden sollten, inwiefern Lernerleistungen ein spezifisches Niveau erreicht haben oder nicht:

> if one needs to determine whether a student has reached one specific level, it is worth exploring an approach in which tasks are used that are each targeted at one specific level; the written responses of the students are then assessed by having trained raters assign a fail/pass rating using level-specific rating instruments (Harsch & Rupp 2011: 2).

Gemäß den beiden Forschern sind die einzelnen Aufgaben des Tests in einem niveauspezifischen Ansatz dem zu überprüfenden Niveau angepasst und mit trainierten Bewertenden können somit zuverlässige Aussagen über dieses Niveau getroffen werden.

Ausgehend von diesen Erkenntnissen wird in der vorliegenden Untersuchung eine Zertifikatsprüfung des schriftlichen Ausdrucks in Deutsch aus dem Goethe-Institut verwendet. Die vorliegende Studie folgt dem niveauspezifischen Ansatz, daher wurde ein Test auf dem Niveau B1 des GER gewählt. Die Wahl fiel auf eine der weit verbreiteten Zertifikatsprüfungen des Goethe-Instituts (*Goethe-Zertifikat B1*). Diese Zertifikatsprüfungen orientieren sich an den Sprachreferenzniveaus des GER und sind durch die Organisation *Association of language testers in Europe* (ALTE) zertifiziert (vgl. Kap. 2.3.2). Bei der Zuordnung der Prüfung *Zertifikat B1* zum B1-Niveau des GER im Jahr 2012 wurden die vorgesehenen Schritte, die im *Manual* (Council of Europe 2009) vertreten sind, befolgt (Glaboniat et al. 2013). Dabei sollte u. a. nachgewiesen werden, dass die Anforderungen der Prüfung dem angestrebten

Niveau entsprachen. Des Weiteren sollte die Bestehensgrenze für die Teilnehmerleistungen, die als bestanden gelten, bestimmt werden und diese Leistungsbeispiele sollten ebenfalls mit dem angestrebten Niveau verglichen werden (ibid.).

Den Zertifikatsprüfungen liegen somit die Qualitätsstandards der ALTE zugrunde und die Qualität wird zudem durch regelmäßige Kontrollen sichergestellt (vgl. Goethe-Institut 2018). Die Prüfungen werden für Teilnehmende ab 16 Jahren von Goethe-Instituten in Deutschland und weltweit für jede Stufe der sechsstufigen Kompetenzskala des GER angeboten. Das Goethe-Zertifikat entspricht demzufolge den Niveaustufen des GER vom Anfänger (A1) bis zum avancierten Sprachverwender (C2) und stellt eine Möglichkeit zur Operationalisierung des GER-Standards dar. Bei diesen Tests wird nur mit ganzen Niveaustufen gearbeitet und eine weitere Unterteilung in einen oberen bzw. unteren Bereich der Stufe (z. B. A2.2 oder B1.1) wird nicht gemacht. Die weltweite Verwendung der Zertifikate A1–C2 des Goethe-Instituts mit 230 000 Prüfungsteilnehmenden pro Jahr (Goethe-Institut 2019), die Qualitätssicherung der Prüfungen entsprechend den GER-Niveaus und der niveauspezifische Ansatz begründen die Anwendung dieser Prüfung, um die Ziele der vorliegenden Arbeit zu erreichen.

Auf dem B1-Niveau besteht die Zertifikatsprüfung aus vier Prüfungsteilen, drei schriftlichen Modulen: 1) Hören, 2) Lesen, 3) Schreiben und einem mündlichen Modul: 4) Sprechen. Der für diese Arbeit relevante Prüfungsteil ist der Test des schriftlichen Ausdrucks. Dieser Test besteht aus drei Aufgaben, die die schriftliche Sprachfertigkeit der Teilnehmenden in unterschiedlichen kommunikativen Kontexten prüfen sollen. Dies entspricht Hinweisen auf die Notwendigkeit von mehr als einer Aufgabe in einem Test, um Varianzen in der Leistung aufgrund möglichen Aufgabeneffekten zu reduzieren. Weir (2005: 69) schreibt hierzu: „The more samples of a student's writing in a test, the more reliable the assessment is likely to be and the more confidently we can generalize from performance on the test tasks". Dies bedeutet, dass die Reliabilität sowohl in Bezug auf die Abdeckung des Inhalts als auch auf die Zuverlässigkeit der Ergebnisse mit mehreren Aufgaben in einem Test erhöht werden kann und dass wir somit verbindlicher die Leistung im Test zum realen Verhalten generalisieren können. Für den Test des schriftlichen Ausdrucks aus der Goethe-Zertifikatsprüfung haben die Teilnehmenden insgesamt 60 Minuten zur Verfügung. In der folgenden Tabelle (vgl. Tab. 10) wird ein Überblick über die Prüfungsziele und Aufgabentypen des Goethe-Zertifikats auf einem B1-Niveau gegeben (Goethe-Institut 2017):

Tab. 10: *Modul Schreiben zur Prüfung Goethe-Zertifikat B1 im Überblick*

Aufgabe	*Prüfungsziel*	*Aufgabentyp*
1	Interaktion Persönliche Mitteilung zur Kontaktpflege	Freies Schreiben (beschreiben, begründen, einen Vorschlag machen)
2	Produktion Persönliche Meinung zu einem Thema äußern	Freies Schreiben (beschreiben, begründen, erläutern, vergleichen, Meinung äußern, usw.)
3	Interaktion Persönliche Mitteilung zur Handlungsregulierung	Freies Schreiben (sich entschuldigen, um etwas bitten, o. Ä.)

Das Prüfungsmaterial der vorliegenden Studie bestand aus einem zur Zeit der Untersuchung noch nicht veröffentlichten Übungssatz für das Modul Schreiben des Goethe-Zertifikats auf B1-Niveau (vgl. Anhang 9). Die Verwendung des Materials wurde der Wissenschaftlerin gestattet und erfolgte mit schriftlicher Zustimmung der Zentrale des Goethe-Instituts in München. Bei der ersten Aufgabe zum Prüfungsteil Schreiben handelte es sich um das Verfassen eines Briefes an eine Freundin/einen Freund. Die folgenden Stichpunkte waren vorgegeben: eine Beschreibung über ein Praktikum in einer Buchhandlung, eine Begründung, was am Praktikum gut war, und ein Vorschlag für ein Treffen. In den Anweisungen stand, dass die erste Aufgabe eine Mindestwortzahl von 80 Wörtern enthalten sollte. Die Lernenden mussten darüber hinaus etwas zu allen drei Inhaltspunkten schreiben und dabei auf den Textaufbau (Anrede, Einleitung, Reihenfolge der Inhaltspunkte und Schluss) achten. In der zweiten Schreibaufgabe geht es um einen Beitrag in einem Online-Forum einer Zeitung zum Thema „private Fotos in sozialen Netzwerken" (80 Wörter), in dem die Lernenden die eigene Meinung zum Thema schreiben sollten. Die dritte Aufgabe bestand aus einer formellen E-Mail. Die Schülerinnen und Schüler sollten darin eine höfliche Entschuldigung mit der dazugehörigen Begründung, warum eine Hausaufgabe nicht gemacht worden ist, an einen Lehrer verfassen (40 Wörter). Der schriftliche Test deckt somit Aktivitäten ab, die in den Sub-Skalen des GER zur schriftlichen Produktion und Interaktion auf einem B1-Niveau beschrieben sind (vgl. Europarat 2001: 67–68; 86–87).

Bewertende

Die Bewertenden der Studie können in drei Gruppen eingeteilt werden: die schwedischen Deutschlehrkräfte der verschiedenen Schülergruppen,

die externen schwedischen Bewertenden und die externen GER-Bewertenden. Für die externe Bewertung wurden jeweils zwei unabhängige Bewertende je externer Bewertergruppe ausgewählt. Auswahlkriterien waren Geschlecht, Alter, Berufserfahrung und geographischer Wohnort. Im Einklang mit den Prinzipien zur Forschungsethik (vgl. Vetenskapsrådet 2002) haben sämtliche teilnehmenden Bewertenden schriftliche Informationen über das Forschungsvorhaben und Kontaktinformationen der Wissenschaftlerin erhalten. Des Weiteren wurde über die Freiwilligkeit der Teilnahme aufgeklärt sowie die Anonymität und der Datenschutz zugesichert. Alle teilnehmenden Bewertenden haben ein Formular mit einer Einverständniserklärung sowie Fragen, u. a. über Alter, Berufserfahrung und Lehrerausbildung ausgefüllt.

Die schwedischen Deutschlehrkräfte der Studie ($N = 18$) waren ausgebildete und praktizierende Lehrerinnen und Lehrer an schwedischen Gymnasien. Die Lehrkräfte wurden per E-Mail kontaktiert, nachdem ein Brief an die Schule geschickt worden war und die Leitung der Schule eine Genehmigung für die Studie gegeben hatte. Der überwiegende Anteil der Lehrkräfte in der Studie hatte jahrelange Berufserfahrung als Lehrkraft. Sie waren im Alter von 31–67 Jahren, zum überwiegenden Teil aber älter als 50 Jahre alt (vgl. Tab. 42, Anhang 10) und hatten eine abgeschlossene pädagogische Ausbildung.[74] Die Bewertung der schriftlichen Leistungen erfolgte gemäß den schwedischen Bildungsstandards, wobei Texte mit den Noten A bis E als bestanden gelten und Texte mit der Note F eine nicht bestandene Leistung bedeuten.

Gemäß Bachman und Palmer (2010) kann ein einzelner Bewertende „have a ‚bad day‘, or be overly lenient or severe in his ratings“ (S. 354). Sie empfehlen daher für eine Einstufung mindestens zwei Bewertungen pro Leistung und eine ergänzende dritte Bewertung, wenn die Testergebnisse der Bewertende weit auseinander liegen (ibid.). Ferner hat Dalberg (2019) in einer Studie zeigen können, dass der zusätzliche Nutzen von mehr als zwei Bewertenden schnell abnimmt. Um Aspekte der Bewerterübereinstimmung untersuchen zu können und die Reliabilität schwedischer Bewertungen von Schülerleistungen für den Vergleich zu einem Referenzniveau des GER zu stärken, konnten zusätzliche unabhängige Bewertende, zwei schwedische Bewertende und zwei GER-Bewertende, für die Studie gewonnen werden.

---

74 Dies spiegelt die Tatsache wider, dass die große Mehrheit der Gymnasiallehrkräfte für Deutsch eine abgeschlossene pädagogische Ausbildung haben und das Durchschnittsalter von Gymnasiallehrkräften für Fremdsprachen in Schweden gemäß der Statistik der schwedischen Schulbehörde ziemlich hoch ist (vgl. Skolverket 2017a).

Die beiden externen schwedischen Bewertenden ($N = 2$), eine weibliche und ein männlicher, waren ausgebildete und erfahrene Gymnasiallehrkräfte im Fach Deutsch (vgl. Tab. 43, Anhang 10). Die Auswahl der externen Bewertenden wurde so vorgenommen, dass unterschiedliche Perspektiven abgebildet werden konnten. Sie repräsentierten unterschiedliche Schulen, Altersgruppen und Regionen in Schweden. Die beiden Bewertenden hatten zudem in ihrem Berufsleben neben ihrer Arbeit als Gymnasiallehrkraft zusätzliche Aufträge im Bereich Bewertung, allerdings unterschiedlicher Art, gehabt. Diese zusätzlichen Erfahrungen in Bezug auf Schülerbewertungen waren ein Grund für die Auswahl der beiden Prüfenden. So ist gewährleistet, dass sie über sowohl kontextuelles Wissen über das schwedische Schulsystem als auch Kenntnisse und Erfahrungen im Bereich Bewertung verfügten. In einer E-Mail haben die externen schwedischen Bewertenden die Informationen zur Studie bekommen. Wie die praktizierenden Lehrkräfte, haben sie eine Bewertung gemäß den schwedischen Bildungsstandards bei der Einstufung der Schülerleistungen vorgenommen, wobei die Noten A bis F vergeben wurden.

Die beiden GER-Bewertenden ($N = 2$), eine weibliche Prüferin und ein männlicher Prüfer, waren beide zertifizierte Bewertende des Goethe-Institutes (vgl. Tab. 44, Anhang 10). Auch sie repräsentierten unterschiedliche Erfahrungen. Die GER-Bewertenden hatten unterschiedlich lange Erfahrung darin, Fremdsprachenkenntnisse zu bewerten. Während der etwas ältere GER-Bewertende eine lange und umfangreiche Erfahrung im Bereich Bewerten von Fremdsprachenkenntnissen hatte (u. a. als Prüfer für *TELC* für Deutsch und als Leiter von Prüferschulungen), konnte der jüngere GER-Bewertende auf weniger Erfahrung zurückgreifen. Es ist in diesem Zusammenhang wichtig zu erwähnen, dass Prüfende der Goethe-Zertifikate seit 2014 eine formale Prüferschulung absolviert haben müssen.[75] Durch eine Vereinbarung über die Unterstützung des Forschungsprojektes wurden diese Bewertenden vom Goethe-Institut in Stockholm beauftragt, die Schülerleistungen zu evaluieren. Bei der Evaluierung der beiden GER-Bewertenden wurde geprüft, inwiefern die Schülertexte die Anforderungen auf dem B1-Niveau für die schriftliche Kompetenz erfüllen.

75 Prüferschulungen sind eine Voraussetzung für Prüferinnen und Prüfer des Goethe-Zertifikats und danach muss eine Schulung pro Prüferin/Prüfer alle drei Jahre abgelegt werden. Diese Zertifizierung der Prüferleistung soll alle fünf Jahre erneuert werden (vgl. Goethe-Institut 2019).

Bewertungsskalen

Die Bewertungsskalen in der vorliegenden Studie basieren auf den schwedischen Bildungsstandards für *Moderna språk* und den Deskriptoren und Skalen des Niveaus B1 im GER. Die Beschreibungen und Anforderungen der schwedischen Lehrpläne für die Fremdsprachen orientieren sich zwar an den GER-Standards; zu bemerken ist aber, dass dennoch keine absolute Übereinstimmung zwischen den jeweiligen Dokumenten herrscht. In der vorliegenden Arbeit ist versucht worden, einen möglichst natürlichen Ablauf bei der Bewertung der Textproduktionen zu gewährleisten, um eine möglichst authentische Bewertung untersuchen zu können. Dies bedeutete, dass die jeweiligen Bewertenden nach ihrem normalen Bewertungsverlauf und nach gewohnten Kriterien die Textproduktionen evaluieren sollten.

Für das Bewertungsverfahren standen entsprechend den schwedischen Lehrkräften und den externen schwedischen Bewertenden die schwedischen Bewertungskriterien der jeweiligen Fremdsprachenstufen zur Verfügung (vgl. Skolverket 2011a). Zusätzliche Informationen bietet den Lehrkräften das Bewertungsmaterial aus der nationalen Prüfungsdatenbank für Fremdsprachen. Dieses Material gibt durch kommentierte und bewertete Schülerbeispiele und aufgestellten Bewertungsfaktoren eine Orientierung, wie die Anforderungen im Lehrplan zu interpretieren sind. Die eher analytisch ausgerichteten Bewertungsfaktoren des nationalen Materials sind in die Dimensionen *Inhalt* und *Sprache und Ausdrucksfähigkeit* eingeteilt (vgl. Anhang 11). Dieses Bewertungsmaterial ist in Deutsch für die Stufen *Tyska 2*, *Tyska 3* und *Tyska 4* zu erhalten, für die niedrigere Stufe 1 und die höheren Stufen 5–7 ist in der Prüfungsdatenbank hinsichtlich der zweiten Fremdsprache kein zusätzliches Prüfungsmaterial vorhanden. Eine weitere Alternative bieten Bewertungschecklisten oder analytische Bewertungsraster, die lokal von Lehrkräften herausgearbeitet worden sind.

Der Referenzrahmen bildet die Grundlage für eine globale Beschreibung der schriftlichen Produktion und Interaktion auf einem B1-Niveau (vgl. Anhang 3 bzw. 4). Die Deskriptoren und Skalen des B1-Niveaus hinsichtlich schriftlicher Interaktion und Produktion sind in einem Bewertungsraster für das Goethe-Zertifikat umgesetzt worden. Die GER-Bewertenden der vorliegenden Studie haben demzufolge die Bewertungskriterien, die zum Prüfungsteil *Schriftlicher Ausdruck* des Goethe-Zertifikats auf dem B1-Niveau gehören, verwendet (siehe Anhang 12). Die im Bewertungsraster hervorgehobenen Bewertungsdimensionen sind *Erfüllung*, *Kohärenz*, *Wortschatz* und *Strukturen*. Zur Dimension *Erfüllung* gehören Aspekte wie Inhalt, Umfang, Textsorte und soziokulturelle

Angemessenheit. Bei der Dimension *Kohärenz* handelt es sich um den Textaufbau und die Verknüpfung von Sätzen. Für sowohl *Wortschatz* als auch *Strukturen* gibt es eine Distinktion zwischen Spektrum (Differenziertheit) und Beherrschung (z. B. Morphologie und Orthographie). Die Bewertungsdimensionen zu Wortschatz und Grammatik stehen in den Kriterien explizit mit der Dimension *Verständnis* in Verbindung.

Das Bewertungsraster umfasste folglich, wie bereits oben erwähnt, vier Bewertungsdimensionen: *Erfüllung*, *Kohärenz*, *Wortschatz* und *Strukturen*. Diese Dimensionen wurden für jede Teilaufgabe auf einer fünfgradigen Skala bewertet und für jede Dimension wurden Punkte vergeben. Insgesamt können in diesem Prüfungsteil maximal 100 Punkte erreicht werden. Diese eher analytisch geprägten Kriterien werden, im Hinblick auf das Erreichen bzw. das Nicht-Erreichen eines B1-Niveaus, zu einer Gesamtbeurteilung zusammengerechnet. Die schriftlichen Schülerleistungen erhielten somit am Ende eine Gesamtpunktzahl, wodurch folgende Prädikate ermittelt wurden: 0–59,5 *nicht bestanden*; 60–69,5 *ausreichend*; 70–79,5 *befriedigend*; 80–89,5 *gut*; 90–100 *sehr gut* (vgl. Goethe-Institut 2018). Eine Schülerleistung der vorliegenden Studie befindet sich dementsprechend auf einem B1-Niveau im Schreiben, wenn insgesamt mindestens 60 Punkte erreicht wurden.

Ablauf bei der Datenerhebung

Die standardisierte Datenerhebung fand im Frühjahr 2017 an Gymnasialschulen in Süd- und Mittelschweden statt. Da die jeweiligen Schulleitungen entscheiden sollten, ob ihre Schule an dem Projekt teilnehmen würde, wurden im März 2017 Briefe an Direktorinnen und Direktoren insgesamt 50 schwedischer Gymnasien, sowohl kommunalen Schulen als auch sog. freien Schulen[76], verschickt. In diesem Brief erhielten die Schulleitungen Informationen über die Studie und eine Anfrage, ob am Ende des Semesters an ihrer Schule Material für die Studie erhoben werden könnte. Mit der Leitung der Schule wurde danach telefonisch Kontakt aufgenommen. Bei einer Zusage der Schulleitung wurden die Deutschlehrkräfte an jener Schule per E-Mail kontaktiert.

76 Eine freie Schule (*friskola*) ist in Schweden eine autonome Schule, die aber wie kommunale Schulen vom Staat finanziert wird. In Schweden gibt es u. a. eine Diskussion darüber, dass freie Schulen eine großzügigere Notengebung als kommunale Schulen pflegen (vgl. Vlachos 2019; Skolverket 2019a; Skolverket 2020b). Dies ist nicht Gegenstand der vorliegenden Studie, kann aber bei der Interpretation der Ergebnisse von Bedeutung sein.

Etwa die Hälfte der Gymnasialschulen hat ihre Teilnahme an der Studie abgelehnt. Der Großteil dieser Entscheidungen beruht auf einer Absage der Lehrkräfte (76 %) im Vergleich zu Absagen seitens der Schulleitung (24 %). Nach einer Ausfallanalyse konnte festgestellt werden, dass ein höherer Anteil derjenigen Gymnasialschulen, die in größeren Städten liegen sowie derjenigen, die als freie Schulen gelten, ihr Mitwirken an der Studie abgelehnt hat. Wenn die Ausfallquote für die Leitung der Schule betrachtet wird, sind es vorwiegend Schulleitungen an kommunalen Gymnasialschulen in größeren Städten, die das Mitwirken ihrer Schule abgelehnt haben. Schaut man sich hingegen die Ausfälle bei den freien Schulen an, zeigt sich zunächst, dass die Deutschlehrkräfte an freien Schulen häufiger als Lehrkräfte kommunaler Schulen ihr Mitwirken an der Studie ablehnen. Diese Lehrkräfte gaben oft an, dass die Schülergruppen im Fach Deutsch an ihren Schulen relativ klein waren und überdies oft auch auf mehrere Stufen verteilt waren, was als eine Erklärung für die Absage angeführt wurde. Unter den Deutschlehrkräften war insgesamt die Mehrheit der Studie gegenüber positiv eingestellt, aber einige Lehrkräfte lehnten ihr Mitwirken dennoch ab, hauptsächlich aufgrund von Zeitmangel am Ende des Semesters. Andere Gründe für Ausfälle waren u. a. Krankschreibungen, Schüleraustausche oder dass die Lehrkraft den Kurs zum ersten Mal unterrichtete.

Insgesamt haben sich 21 Schulen bereit erklärt, an der Datenerhebung teilzunehmen. Zwei Schulen sagten aber spät ab; eine freie Schule aufgrund von Zeitmangel bzw. ein kommunales Gymnasium wegen einer Krankschreibung der Lehrkraft. An der Datenerhebung beteiligt waren dementsprechend 19 Schulen (Teilnahmequote 38 %) und damit insgesamt 25 Schülergruppen. Eine dieser 19 Schulen musste nachträglich ausgeschlossen werden, weil erst bei der Datenerhebung festgestellt wurde, dass die Probanden dieser Schule am Ende des Schuljahres nur die Hälfte des Kurses belegt hatten. Es handelte sich dabei um eine Schülergruppe auf *Tyska 4*. Es hat sich aber erwiesen, dass dennoch genug Schülerinnen und Schüler dieser Stufe in der Studie teilgenommen haben.

Für das Mitwirken an der Studie konnten sowohl kommunale Schulen als auch freie Schulen gewonnen werden. Die Schulform hat zwar für die Studie selbst nur eine geringe Bedeutung, könnte aber für die Repräsentativität der Studie wichtig sein. Die Schulen wurden unter Berücksichtigung von größeren Städten und kleineren Orten ausgewählt, um eine Streuung der Probanden zu erhalten.[77] Alle Schulen haben das Fach *Tyska* angeboten. Da der Kurs *Tyska 5*

---

77 Von den achtzehn an der Studie beteiligten Schulen lagen acht Gymnasialschulen in größeren Städten und zehn in kleineren Städten und Orten. Insgesamt vier der beteiligten Schulen sind freie Schulen, während es sich bei den anderen vierzehn

nicht an allen Schulen angeboten wurde, wurden darüber hinaus auch Schulen mit sprachlichem Schwerpunkt ausgesucht, um Prüfungsteilnehmende dieses Kurses zu finden. Dieses Verfahren ist ein Beispiel für ein sog. *purposive sampling*, eine Vorgehensweise, die einem Forscher ermöglicht, einen spezifischen Bedarf in einer Studie zu erfüllen (Robson & McCartan 2016). Zusammenfassend kann festgestellt werden, dass die Auswahl der Schulen in der Studie keine Zufallsstichprobe ist, sondern aufgrund der freiwilligen Teilnahme der Schulen und der Lehrkräfte am ehesten einer Gelegenheitsstichprobe entspricht. Obwohl es sich in der vorliegenden Studie also um eine Gelegenheitsstichprobe handelt, ist die Varianz der Schulformen, Schul- und Klassengröße sowie der Schulamtsbezirke sehr hoch. Dennoch muss aber die Nicht-Repräsentativität der Stichprobe bei der Interpretation der Ergebnisse beachtet werden.

Die Schülertexte wurden im Frühjahr 2017 im Zeitraum von Anfang April bis Ende Mai an Gymnasialschulen in Süd- und Mittelschweden unter realistischen Prüfungsbedingungen erhoben. Alle Probanden schrieben denselben schriftlichen Sprachtest unter Aufsicht, wobei sie 60 Minuten Zeit hatten, um drei Schreibaufgaben zu bewältigen. Die schriftlichen Aufgaben wurden danach wieder eingesammelt und die Lehrkraft wurde darüber informiert, dass der Test nicht wieder an die Probanden verteilt werden dürfe, da der Inhalt für kommende Gruppen nicht bekannt gemacht werden sollte. Sicherheitshalber wurden die jeweiligen Probanden in einem Fragebogen vor dem Schreiben des Tests danach gefragt, ob sie diesen Test bereits im Voraus gesehen hatten, was alle verneinten. Die Aufgaben gaben an keiner Stelle Informationen darüber, dass der Test ein B1-Niveau des GER prüfte oder dass dieser vom Goethe-Institut stammte.

Zur Auswahl standen in der Studie insgesamt 225 Schülertexte aus 24 unterschiedlichen Schülergruppen und 18 Schulen. Diese Schülergruppen waren unterschiedlich groß; im Durchschnitt gab es eine höhere Anzahl an Probanden in den Schülergruppen der Kurse *Tyska 3* und *Tyska 4* im Vergleich zu den Schülergruppen des Kurses *Tyska 5*. Aus diesem Grund stammten die

---

um kommunale Schule handelt. Dies bedeutet, dass 22 % der Schulen im Material freie Schulen sind. Im Hinblick auf die Repräsentativität der Daten ist dies in etwa im Einklang mit dem Anteil der schwedischen Probanden, die ihre Gymnasialausbildung im Schuljahr 2016/17 an freien Gymnasialschulen absolvierten, nämlich 25 % (vgl. Skolverket 2017b). Da die Schülergruppen an den freien Schulen in der Studie zahlenmäßig geringer sind als die Schülergruppen an kommunalen Gymnasialschulen, beläuft sich der Anteil der Texte, der aus freien Gymnasialschulen stammt, auf 18 % des gesamten Materials nach Textauswahl.

Schülerleistungen auf *Tyska 3* und *Tyska 4* aus sechs bzw. sieben Schülergruppen, während elf Schülergruppen auf *Tyska 5* teilnahmen, um genügend Texte für die Studie zu erhalten. Die Schülerleistungen aus den Fremdsprachenstufen *Tyska 3*, *Tyska 4* und *Tyska 5* verteilen sie sich auf die einzelnen Noten wie folgt:

Tab. 11: *Verteilung der schriftlichen Schülerleistungen nach Kurs und Note*

Kurs/Note	*F*	*E*	*D*	*C*	*B*	*A*	***Gesamt***
Tyska 3	15	22	10	23	16	4	**90**
Tyska 4	10	14	12	24	13	6	**79**
Tyska 5	6	2	7	14	14	13	**56**
Gesamt	**31**	**38**	**29**	**61**	**42**	**23**	**225**

Wie aus Tab. 11 ersichtlich, erreicht ein größerer Anteil der Leistungen auf *Tyska 5* eine höhere Note als auf *Tyska 3* und *Tyska 4*. Umgekehrt erhält ein größerer Anteil der Schülertexte auf *Tyska 3* eine niedrigere Note E. Dies ist nicht überraschend, da in der Regel Schülerinnen und Schüler mit einer höheren Note ihre gewählte Sprache weiterlernen. Ein anderer Grund für diesen Unterschied liegt aber auch in der Aufgabenstellung des Tests, die für Schülerinnen und Schüler in der Fremdsprachenstufe *Tyska 3* als schwierig empfunden werden könnte, da sie sich eventuell noch nicht auf dem im Test zu prüfenden sprachlichen Niveau befinden.

Textauswahl

Für die vorliegende Untersuchung wurden durch ein Auswahlverfahren insgesamt 60 schriftliche Textproduktionen ausgesucht. Zunächst erfolgte eine systematische Auswahl. Insgesamt sollten gleich viele Texte aus jeder Kursstufe enthalten sein, d. h. jeweils 20 Texte aus den Kursen *Tyska 3*, *Tyska 4* und *Tyska 5*. Die Einstufung der Schülertexte liegt der Auswahl dieser 60 Texte zugrunde. Um Schülertexte mit verschiedenen Notenstufen zu erhalten, wurden 60 Textproduktionen mit möglichst unterschiedlichen Noten, die von den an den Gymnasien unterrichtenden Lehrkräften vergeben worden waren. Da in den schwedischen Lehrplänen explizit Kriterien für die Notenstufen E, C und A vorhanden sind, wurden Textproduktionen mit diesen Einstufungen zusammen mit der nicht ausreichenden Note F gegenüber Texten mit den Noten D und B bevorzugt: für jeden Kurs sollten somit jeweils fünf Schülertexte mit den Noten A, C, E und F ausgesucht werden. In den Fällen, in denen es nicht genug Texte gab, wurden Texte mit ähnlichen Noten genommen. Dies betraf

die Auswahl von Texten mit der Note A auf *Tyska 3*, wobei eine Schülerleistung mit der Note B ausgewählt wurde, und die Auswahl von Textproduktionen mit der Note E auf *Tyska 5*, wobei drei Texte mit der Note D herausgesucht wurden. Die Auswahl der Schülertexte je Note wird in Tab. 12 abgebildet:

Tab. 12: *Verteilung der 60 Schülerleistungen nach Kurs und Note nach dem Auswahlverfahren*

Kurs/Note	*F*	*E*	*D*	*C*	*B*	*A*	***Gesamt***
Tyska 3	5	5		5	1	4	**20**
Tyska 4	5	5		5		5	**20**
Tyska 5	5	2	3	5		5	**20**
Gesamt	**15**	**12**	**3**	**15**	**1**	**14**	**60**

Bei der Auswahl der Texte wurde dem Prinzip gefolgt, möglichst viele unterschiedliche Schulen auszuwählen, um die Generalisierbarkeit der Studie zu erhöhen und um möglichst viele unterschiedliche Bewertende einzubeziehen, damit die Bewertung einer einzelnen Lehrkraft keinen allzu großen Einfluss in der Studie bekommen sollte. Zugleich wurden jedoch auch die Repräsentativität und die Proportionalität des gesamten Materials beachtet. Nachdem die Texte zunächst in Bezug auf Noten und Schule ausgewertet worden waren, erfolgte eine repräsentative proportional stratifizierte Auswahl der Texte, um den Anteil der Unterrichtsgruppen im Gesamtmaterial zahlenmäßig angemessen zu repräsentieren. Bei der abschließenden Auswahl innerhalb der Unterrichtsgruppen wurde eine Zufallsauswahl angestrebt.

Die handschriftlichen Texte wurden zunächst transkribiert, um zu vermeiden, dass eventuelle Korrekturen der Deutschlehrkräfte die externen Bewertenden beeinflussten. Bisherige Kommentare anderer Bewertender können sonst einen erheblichen Einfluss auf die Bewertung haben (vgl. Murphy 1979). Allerdings haben frühere Studien gezeigt, dass handgeschriebenen Texten im Durchschnitt höhere Punktzahlen als computerbasierten Texten verliehen wurden, unabhängig davon, ob der Text per Hand oder am Computer geschrieben war (Powers et al. 1994). Gründe dafür könnten sein, dass die computergeschriebenen Texte im Vergleich mit der handgeschriebenen Version weniger umfangreich erscheinen und dass einige Fehler wie Falschschreibungen und grammatische Fehlentscheidungen in computerproduzierten Texten sichtbarer sind. Die transkribierten Texte der vorliegenden Studie waren aus diesem Grund doppelzeilig, damit sie nicht als kurz aufgefasst werden sollten und wurden per Brief an die Bewertenden geschickt, um eventuelle digitale

Schreibkontrollen zu vermeiden. Die Tatsache, dass die praktizierenden Lehrkräfte handgeschriebene Schülertexte evaluierten, während die externen Bewertenden transkribierte Schülertexte erhielten, könnte trotzdem bei der Bewertung einen Einfluss gehabt haben.

Bewertungsprozess und Übersicht der schriftlichen Kommentare

Das Ziel war es, die Bewertungssituation so authentisch wie möglich zu gestalten. Die teilnehmenden Deutschlehrkräfte haben daher kein Bewertertraining oder andere Anweisungen erhalten und sollten so verfahren, wie sie normalerweise schriftliche Schülerleistungen auf Deutsch bewerten. Bei der Bewertung konnten sie zusätzlich zu den Anweisungen der schriftlichen fakultativen Tests aus der Prüfungsdatenbank die dazugehörigen Benchmark-Beispiele für *Tyska 3* und *Tyska 4* als Unterstützung nutzen. Darüber hinaus sind beim Evaluieren von Schülertexten oft schriftliche Kommentare oder Begründungen zur gegebenen Note inbegriffen, eine Begründung der Note zu geben gehört somit für die Lehrkräfte zum üblichen Verfahren. Viele der teilnehmenden Lehrkräfte gaben auch an, dass sie bei der Beurteilung ihrem normalen Prozedere gefolgt waren. Die Schülerleistungen wurden folglich in einem ersten Schritt von der jeweiligen Lehrkraft (N = 18), die die Prüfungsteilnehmenden unterrichtet hatte, gemäß den schwedischen Bewertungskriterien auf einer sechsgradigen Skala mit den Noten F bis A evaluiert.

Hierauf folgte nach dem oben beschriebenen Auswahlprozess eine Bewertung von 60 ausgewählten Schülerleistungen durch die zwei externen, voneinander unabhängigen, schwedischen Bewertenden. Sowohl die Gruppe der Lehrkräfte als auch die externen schwedischen Bewertenden bewerteten die Textproduktionen gemäß den schwedischen Kriterien und gaben zusätzlich eine schriftliche Begründung für die Evaluation.

Anschließend erteilte das Goethe-Institut Schweden durch eine Vereinbarung über die Unterstützung des Forschungsprojektes zwei für die Goethe-Zertifikate ausgebildeten Prüfenden den Auftrag, die schriftlichen Prüfungsteile der Probanden getrennt zu bewerten. Diese GER-Bewertenden evaluierten jede Schülerleistung anhand von Kriterien, die auf dem GER basieren. Hierbei wurde geprüft, inwiefern die Texte die Kriterien für ein B1-Niveau erfüllten bzw. nicht erfüllten. Die Evaluation der GER-Prüfenden erfolgte nach dem vom Goethe-Institut festgelegten Bewertungsraster (vgl. Anhang 12).

Bei der Bewertung haben sämtliche Bewertende neben einer Note für jede Schülerleistung zusätzlich auch eine schriftliche Begründung der Note gegeben. Das Datenmaterial besteht folglich aus einer quantitativen Komponente, den Testergebnissen (die Noten F–A und ein erreichtes oder nicht-erreichtes Niveau

B1 des GER), sowie aus einem qualitativen Teil mit schriftlichen Kommentaren der Bewertenden. Auch in bisherigen Studien wurden Begründungen der Leistungsbeurteilung durch die Bewertenden verwendet (z. B. Brown et al. 2005; Barkaoui 2010a; Borger 2018). Hier wurde aber auf sog. TAP-Methoden (*Think-aloud-protocols*), wo Bewertende ihre Gedanken während der Beurteilung verbalisieren, verzichtet. Es besteht die Gefahr, dass TAP-Methoden Auswirkungen auf den Bewertungsprozess haben könnten und sie sollten aus diesem Grund in Studien, wo der Fokus eher auf der Bewertung liegt, vermieden werden (vgl. Barkaoui 2011b). Ein solches Verfahren hätte dementsprechend das normale Prozedere der Bewertenden beeinflussen können und daher wurden schriftliche Bewerterkommentare für die vorliegende Studie bevorzugt.

Das qualitative Material ist aus insgesamt 300 geschriebenen Kommentaren zusammengesetzt, wovon 60 von unterschiedlichen schwedischen Lehrkräften, 120 von den beiden externen schwedischen Bewertenden und 120 von den beiden GER-Bewertenden verfasst wurden. Jeder der 60 Schülertexte wurde demzufolge von der unterrichtenden Lehrkraft, zwei externen schwedischen Bewertenden und zwei externen GER-Bewertenden beurteilt (vgl. Anhang 13). Somit sind zu jedem Text fünf geschriebene Begründungen für die Bewertung vorhanden. Alle schriftlichen Kommentare im Text, sowohl vereinzelte Anmerkungen am Textrand als auch zusammenhängende Erläuterungen, wurden transkribiert. Korrekturen der Bewertenden in den Schülertexten wie ein Plus, ein Fragezeichen oder eine Unterstreichung wurden bei dieser Analyse jedoch nicht berücksichtigt, da nicht immer eindeutig war, was die/der Bewertende damit gemeint hat. Die Kommentare der Bewertenden sehen überdies unterschiedlich aus. Die Bewertenden, die den schwedischen Kriterien gefolgt sind, variieren stark in Bezug auf sowohl die Textmenge als auch die Art der Kommentare. Zwei der Deutschlehrkräfte verwendeten bei der Bewertung ihrer Schülertexte ein lokales Bewertungsraster und markierten in diesem die Kriterien, die die Schülerinnen und Schüler in den Texten erfüllt hatten. Wenn ein Bewertungsraster verwendet wurde, wurden nur die Teile in die Analyse miteinbezogen, die die Lehrkraft im Hinblick auf den zu bewertenden Text im Bewertungsraster markiert hatte.

Die GER-Bewertenden verwendeten bei der Beurteilung ein Bewertungsraster für schriftliche Leistungen für ein B1-Niveau des Goethe-Instituts, wobei dieses Bewertungsraster, das sich auf die vier Hauptdimensionen (*Struktur*, *Wortschatz*, *Erfüllung* und *Kohärenz*) bezieht, die Grundlage ihrer schriftlichen Kommentare bildete (vgl. Anhang 12). Dies bedeutet aber auch, dass die GER-Bewertenden eine eher analytisch geprägte Bewertung durchgeführt haben, während die Bewertenden, die den schwedischen Kriterien gefolgt sind, häufiger eine eher holistisch geprägte Bewertung durchgeführt haben.

5.3 Analyseverfahren

In diesem Kapitel wird auf die Auswertungsmethoden der gesammelten Daten eingegangen. Mit Hinblick auf die Beantwortung der drei Forschungsfragen gliedert sich das Analyseverfahren in zwei Teile, einen qualitativen und einen quantitativen Ansatz. Die Daten dieser Mixed-Methods-Studie bestehen aus zwei Teilen: den Ergebnissen der Schülerleistungen und den schriftlichen Kommentaren der Bewertenden. Während die Testergebnisse der Schülerleistungen in Form einer Punktzahl oder einer Note durch quantitativen Methoden berechnet wurden, wurden die Kommentare der Bewertenden vorwiegend nach qualitativen Methoden untersucht. Die Auswertungsmethoden lassen sich im Hinblick auf die Forschungsfragen der vorliegenden Arbeit (vgl. Kap. 1.1) in Tab. 13 zusammenfassen:

Tab. 13: *Überblick über die qualitativen bzw. quantitativen Auswertungsmethoden*

Fragestellung	*Daten*	*Qualitativer Ansatz*	*Quantitativer Ansatz*
1	Schriftliche Kommentare der schwedischen Bewertenden bzw. GER-Bewertenden	Deduktiv und induktiv basierte thematische Inhaltsanalyse	Prozentuale Berechnungen der Kodierkategorien
2	Testergebnisse (Noten F–A) sowie schriftliche Kommentare der schwedischen Bewertenden	Qualitative Vergleiche der schriftlichen Kommentare bei ähnlicher bzw. unterschiedlicher Bewertung der schwedischen Bewertenden	Deskriptive Statistik, Analyse der Bewerterübereinstimmung
3	Schriftliche Kommentare der schwedischen Bewertenden bzw. GER-Bewertenden sowie Testergebnisse (Noten F–A bzw. ein erreichtes oder nicht-erreichtes B1-Niveau)	Qualitative Vergleiche der schriftlichen Kommentare bei Textproduktionen auf *Tyska 5*, die Mindestanforderungen für ein Niveau B1 erfüllen, in Relation zu einer ausreichenden Note E	Deskriptive Statistik, Korrelationsberechnungen

Im Folgenden soll das Analyseverfahren der qualitativen bzw. quantitativen Methoden näher beschrieben werden. Da die Methodik im Hinblick auf die Inhaltsanalyse der Bewerterkommentare aus mehreren unterschiedlichen Phasen besteht, wird dieser Teil ausführlicher behandelt.

5.3.1 Qualitative Inhaltsanalyse

Um die erste Forschungsfrage, d. h. inwiefern zentrale Aspekte auf der Ebene der Texte für die Bewertungen besonders relevant erscheinen, beantworten zu können, wurde eine qualitative Inhaltsanalyse (vgl. Kuckartz 2014b) der schriftlichen Kommentare der Bewertungen vorgenommen. Für diese Untersuchung wurden Kategorien sowohl induktiv als auch deduktiv erzeugt und anhand einer qualitativen Analyse wurden die Bewerterurteile in Segmente aufgeteilt. Diese wurden weiterhin in die erstellten Kategorien und gegebenenfalls in Unterkategorien eingestuft (eine nähere Beschreibung des qualitativen Teils der Inhaltsanalyse folgt unten). Hierbei wurden die Segmente pro Kategorie und nach den jeweiligen Bewertenden in Tabellen erfasst und prozentual berechnet (der quantitative Teil der qualitativen Inhaltsanalyse). Die jeweiligen schriftlichen Kommentare der Bewertenden wurden auch qualitativ analysiert, um zusätzliche Kenntnisse über die den Noten zugrundeliegenden Faktoren erlangen zu können. Nachstehend wird auf das Vorgehen der qualitativen Inhaltsanalyse eingegangen.

Kodierschema und Kodierverfahren: Ausarbeitung und Validierung

Zunächst sollte ein Kodierungsschema realisiert werden. Bei dessen Ausarbeitung wurde die Vorgehensweise von Green (1998) verfolgt. Green (S. 68) warnt davor, dass unterschiedliche Forscher unabhängig voneinander unterschiedliche Kodierschemata entwickeln könnten, wenn sie das gleiche Material kodieren. Laut Green macht dies die Methode nicht ungültig; bei der Interpretation der Ergebnisse sollte aber darauf geachtet werden. Beim Identifizieren der Kategorien wurden zudem die Empfehlungen von Green berücksichtigt, die ein balanciertes Vermeiden von zu vielen oder zu breiten Kategorien vorschlagen:

> A good coding scheme achieves a balance between specificity and generalisability. Poor coding schemes can be either too general, failing to capture adequately the cognitive activity involved in carrying out a task, or too idiosyncratic and thereby failing to represent typical behaviour. (Green 1998: 70–71)

Zu viele spezifische Kategorien können darüber hinaus zu einer geringeren Reliabilität der Analyse führen. Zunächst wurde ein Kodierungsschema entwickelt, das sowohl auf den Bewertungskriterien des GER und den schwedischen Rahmenplänen als auch auf den schriftlichen Kommentaren der Bewertenden basierte. Die Kategorien können sowohl deduktiv auf theoretischen Grundlagen als auch induktiv am Material erarbeitet werden (vgl. Kuckartz 2014a). Die Kategorienbildung erfolgt in den meisten Fällen durch ein gemischtes deduktiv-induktives Verfahren, das mehrere Schritte durchläuft (vgl. Kuckartz 2014b: 70), ein Verfahren, das auch in der vorliegenden Arbeit verfolgt wurde.

In einer ersten Phase wurde das Material thematisch theoriegeleitet (vgl. Kap. 3.1.2) unterschiedlichen Kategorien zugeordnet, wie *Wortschatz* und *soziokulturelle Angemessenheit*. Dabei wurden die Kriterien des GER und die Kriterien der schwedischen Bildungsstandards für Fremdsprachen, aber auch Ansätze bisheriger Studien (vgl. Kap. 4) berücksichtigt. Wie in früheren Studien über relevante Aspekte bei der Beurteilung sprachlicher Kompetenz (vgl. Kim 2009; Barkaoui 2010a) erhielten beispielsweise Begründungen, die eine *pauschale Bewertung* über die Sprachverwendung im Schülertext abgaben, eine eigene Kategorie. Zusätzlich erfolgte auch eine Orientierung an bereits etablierten Leitlinien sprachlicher Kompetenz, z. B. an den in den USA gängigen *ACTFL Proficency Guidelines* des American Council on the Teaching of Foreign Languages (2012). In einer zweiten Phase wurden die Kategorien unter Einbezug der Kommentare der Bewertenden weiterentwickelt und verfeinert. Es handelte sich um Aspekte, die in den schon vorhandenen Kategorien nicht abgedeckt worden waren, wodurch neue Hauptkategorien, wie *Gesamteindruck*, und neue Subkategorien, wie *Textlänge*, im Material generiert wurden. Somit konnten Kategorien sowohl auf der Basis von theoretischen Mustern, Bewertungskriterien und früheren Studien (deduktiver Ansatz) als auch mit dem Material als Grundlage (induktiver Ansatz) gebildet werden. Die Kategorienbenennung bezieht sich folglich auf eine Auswahl von Bewertungsdimensionen, die in den theoretischen Grundlagen oder im Material zu finden sind.

Bei der Ausarbeitung eines Kodierschemas ist es notwendig, dass Kriterien zur Reliabilität, Validität und Objektivität beachtet werden. Bei einer Reliabilitätsprüfung der Kodierung wird kontrolliert, ob mindestens zwei unabhängige Kodierer zu den gleichen oder ähnlichen Ergebnissen wie der Forscher kommen (vgl. Green 1998). Sowohl die Ausarbeitung eines Kodierschemas als auch die Reliabilitätsprüfung wurden mithilfe der beiden voneinander unabhängigen Kodierpersonen durchgeführt. Die Kodierer waren an zwei unterschiedlichen Universitäten als Sprachwissenschaftler mit Erfahrungen im didaktischen Bereich tätig. Insgesamt sollten 30 % des Materials in zwei Etappen kontrolliert werden. Die Kodierer haben in einem ersten Schritt insgesamt zehn Kommentare der unterschiedlichen Bewertenden kodiert. Die Kommentare wurden zufällig ausgewählt, jedoch nach einem systematischen Analyseverfahren: aus jeder Bewertergruppe und innerhalb der Fremdsprachstufen *Tyska 3*, *Tyska 4* und *Tyska 5* verteilt. Daraus folgte eine Diskussion über die Angemessenheit des Kodierschemas, die zu geringfügigen Veränderungen führte, sowie über die Analyseprinzipien. Hierbei wurden zudem die Kodierregeln und die Kategorienbildung diskutiert und festgelegt. Nach der Festlegung der Kategorien besteht das Kodierschema aus zehn Hauptkategorien und insgesamt vierzehn Subkategorien mit dazugehörigen Ankerbeispielen (vgl. Tab. 14):

Tab. 14: *Hauptkategorien, Subkategorien und Ankerbeispiele des Kodierschemas*

Hauptkategorien	*Subkategorien*	*Ankerbeispiele*
A: Gesamteindruck		Alla tre uppgifter har eleven behandlat väl och utförligt.[a]
B: Formale Strukturen	1. Grammatische Korrektheit	Gör verb- och genusfel.[b]
	2. Spektrum	Die Sätze beginnen fast immer mit dem Subjekt.
	3. Orthographie	Achtung: Substantive = groß, Kommas!
C: Wortschatz	1. Wortschatzbeherrschung	Eleven gör inte många ordvalsfel.[c]
	2. Wortschatzspektrum	Grundläggande ordförråd finn*s*.[d]
	3. Idiomatische Ausdrücke	Bra idiomatiska vändningar:[e] „Es nervt mich".
D: Pauschale Beurt. – Sprache		Dock många språkliga fel.[f]
E: Textfluss (*fluency*)		Eleven har ett mycket bra flyt i språket[g].
F: Kommunikative Strategien		Strategier för att få fram vad han vill säga[h].
G: Verständlichkeit	1. Allgemein	Svårförståeligt[i].
	2. Die Verwendung von L1/ Englisch	med en del svenska ord[j]
H: Aufgabenerfüllung (*task fulfilment*)	1. Erfüllung vom Inhalt	Fattas en del av uppgiften för innehållet.[k]
	2. Textlänge	Etwas kurz.
I: Angemessenheit	1. Textaufbau	Avslut saknas![l]
	2. Textsorte	Ok anpassat till texttyp[m]
	3. Kohärenz	Kohärenz: Überwiegend angemessen.
	4. Soziokulturelle Angemessenheit	Informelle Anrede!
J: Sonstiges		En del kreativa påhitt i brevet.[n]

[a] „Alle drei Aufgaben hat der Schüler/die Schülerin gut und ausführlich behandelt". (*Hier und im Folgenden eigene Übersetzung, M.H.R.*).
[b] „Macht Verb- und Genusfehler".
[c] „Der Schüler/die Schülerin macht Fehler in der Wortwahl".
[d] „Grundlegender Wortschatz vorhanden".
[e] „Gute idiomatische Wendungen".
[f] „Jedoch viele sprachliche Fehler".
[g] „Der Schüler/die Schülerin hat eine sehr gute Sprachflüssigkeit".
[h] „Strategien, um das, was er sagen will, verständlich zu machen".
[i] „Schwer verständlich".
[j] „mit einigen schwedischen Wörtern".
[k] „Ein Teil der Aufgabe für den Inhalt nicht vorhanden".
[l] „Abschluss fehlt".
[m] „Der Textsorte angemessen".
[n] „Einige kreative Einfälle im Brief".

Ergänzt wurden die Haupt- und Subkategorien durch sog. Ankerbeispiele, d. h. Zitate, aus dem Material, die intersubjektiv überprüft und kontrolliert worden waren. Ein Verfahren mit Ankerbeispielen kann für Kodierpersonen eine Orientierung sein und diese können bei der Zuordnung der Segmente im Material als Musterbeispiele dienen. Die Ankerzitate aus dem Material sollten entsprechend sowohl die Kategorien veranschaulichen als auch eine eindeutige Zuordnung erleichtern und somit die Validität und die Reliabilität der Kordierung erhöhen (vgl. Mayring 2015).

In einem zweiten Schritt wurde durch ein weiteres Kodierungsverfahren die Interkodier-Übereinstimmung kontrolliert. Es handelte sich dabei um 80 Bewerterkommentare, die sowohl von zwei Kodierpersonen als auch von der Wissenschaftlerin unabhängig voneinander kodiert wurden.[78] Nach individuellen Gesprächen über einige Inkonsequenzen bei der Kodierung wurden die prozentualen Übereinstimmungen (PÜ) für die Zuordnung in Hauptkategorien berechnet. Die prozentualen Übereinstimmungen in der vorliegenden Arbeit wurden paarweise ermittelt, indem überprüft wurde, inwieweit jede einzelne Kodierperson mit jeder anderen Kodierperson übereinstimmt. Eine paarweise gemittelte Übereinstimmung ist gegenüber einer gesamten prozentualen Übereinstimmung zu bevorzugen, damit ein Urteil einer Kodierperson nicht allzu starkes Gewicht bekommt (vgl. Wirtz & Caspar 2002: 49).

Für den zweiten Interkodier-Durchlauf beträgt hier die prozentuale Übereinstimmung zwischen den beiden Kodierern 86,3 %, zwischen Kodierperson 1 und der Wissenschaftlerin 90,1 % sowie zwischen Kodierperson 2 und der Wissenschaftlerin 94,1 %. Damit weisen die Werte auf eine nach Wirtz & Caspar (2002) als zufriedenstellend anzusehende Übereinstimmung hin. Die Werte zeigen aber gleichzeitig auch wie komplex das Kodieren, trotz intensiven Diskussionen, manchmal sein kann. Danach wurden alle restlichen Bewerterkommentare von der Wissenschaftlerin Haupt- und Unterkategorien zugeordnet. Das gesamte Kodierungsvorgehen dient als Qualitätskontrolle und soll die Reliabilität der Kategorienbildung und der Kodierung stärken. Nach der Kodierung in Haupt- und Unterkategorien wurden die Frequenzen und Proportionen der Kategorien bezüglich der jeweiligen Bewertenden sowie der unterschiedlichen Bewertergruppen zusammengestellt und berechnet.

---

78 Die erste Kodierperson hat aber aus Zeitgründen lediglich 46 der 80 Bewerterkommentare kodiert. Demzufolge ist die gemittelte paarweise Übereinstimmung bezüglich der Kodierungen für die Kodierperson 1 auf 46 statt auf 80 Kommentare berechnet.

Analyseprinzipien

Die Kommentare wurden in Segmente eingeteilt, die, gemäß Green (1998), immer einen Prozess oder eine Idee repräsentieren sollten. Die unterschiedlichen Segmente repräsentierten somit jeweils einen Aspekt oder eine Idee der Bewertenden bzgl. der Begründung, warum ein Schülertext eine spezifische Note bekommen hatte. Da es schwierig war, abzugrenzen oder zu identifizieren, ob ein Segment einen Aspekt in der Beurteilung wiederholte oder ob es einen neuen Aspekt in der Schülerleistung widerspiegeln sollte, wurden Wiederholungen, Ausarbeitungen oder mehrere Äußerungen zu einem Aspekt in Übereinstimmung mit früheren Studien (vgl. Brown et al. 2005; Ducasse & Brown 2009) immer nur als ein Segment kodiert, was in Beispiel 5.1 ersichtlich ist:

(5.1) Text 1: nicht korrekte Textsorte […] Form = Mail →aber das ist keine Mail (*Angemessenheit – Textsorte*). (Srrs2-3-C[79], Lehrkraft[80])

In Beispiel 5.1 beziehen sich beide Äußerungen auf die Textsorte und werden demnach auch als ein Segment identifiziert. Der Bewertende weist sowohl am Anfang als auch am Ende seines schriftlichen Kommentars auf die Textsorte hin und wiederholt damit den gleichen Aspekt als Begründung seiner Beurteilung.

Darüber hinaus kam es dennoch vor, dass eine Textpassage mehrere Bedeutungen enthielt. Dies war z. B. bei Aufzählungen der Fall, die mehrere Aspekte beinhalteten und bei der Kodierung mehreren Kategorien zugeordnet worden waren, siehe Beispiel 5.2:

(5.2) Mycket korta (*Aufgabenerfüllung – Textlänge*), inte helt begripliga svar (*Verständlichkeit – allgemein*) med en del svenska ord (*Verständlichkeit – die Verwendung von L1/Englisch*).[81] (Imns4-3-F, Lehrkraft)

Im obigen Beispiel wurden die Kommentare in mehrere Teile gegliedert; der erste Teil wurde als Hauptkategorie *Aufgabenerfüllung* aufgeführt und der zweite Teil als *Verständlichkeit*. In Beispiel 5.2 wurden auch die Unterkategorien

---

79 Bezieht sich auf die anonymisierte Buchstaben-Zahlenkennung, auf die Fremdsprachenstufe jeder einzelnen Textproduktion sowie auf die Benotung der jeweiligen schwedischen Bewertenden.

80 Bezieht sich auf die unterschiedlichen Bewertenden: die Gruppe der Lehrkräfte (*Lehrkraft*), die externen schwedischen Bewertenden (*ext. schwed. Bewert. 1* und *ext. schwed. Bewert. 2*) sowie die externen GER-Bewertenden (*GER-Bewert. 1* und *GER-Bewert. 2*).

81 „Sehr kurze, nicht ganz verständliche Antwort mit einigen schwedischen Wörtern". (*Hier und im Folgenden eigene Übersetzung, M. H. R.*).

angegeben. Bei der Aufgabenerfüllung wurde der Kommentar in die Unterkategorie *Textlänge* eingestuft und bei der Verständlichkeit in die Unterkategorien *Verständlichkeit – allgemein* bzw. *Verständlichkeit – die Verwendung von L1/Englisch*.

Bei der qualitativen Inhaltsanalyse wurden sämtliche Kommentare berücksichtigt und in die unterschiedlichen Kategorien eingeordnet. Wie aus den Ankerbeispielen ersichtlich wird, waren die Begründungen zu den Bewertungsdimensionen entweder positiv, neutral oder negativ formuliert. Durch eine Analyse, ob die Bewertenden unterschiedliche Aspekte verschiedener evaluativer Kommentare abgedeckt hatten, konnten relevante Informationen über den Bewertungsprozess sichtbar gemacht werden. Aufbauend auf diesen Überlegungen und in Anlehnung an frühere Studien (vgl. Vaughan 1991; Rinnert & Kobayashi 2001; Barkaoui 2010a; May 2011; Borger 2018) wurde daher untersucht, wie hoch der Anteil positiver, neutraler und negativer Formulierungen war, verteilt auf die unterschiedlichen Kategorien unter den Kommentaren.

Verkettung von Daten durch QDA-Computer Software

Die Organisation und Analyse der Daten erfolgte mithilfe der QDA-Software NVivo 12.[82] Die Textpassagen wurden hierbei manuell markiert und den passenden Kodierkategorien zugeordnet. Anschließend wurden alle mit den Kategorien kodierten Textstellen mit den jeweiligen Bewertenden verbunden. Dadurch entstand eine komplexe Baumstruktur, die zusätzliche Relationen veranschaulichen kann. Der Einsatz der QDA-Software ermöglicht und unterstützt eine kategorienbasierte Auswertung der Bewerterkommentare, bietet aber keineswegs eine vollständige Methodik. Die inhaltsanalytischen Prozeduren wie die Analyseeinheiten, die Kategoriendefinition und der Kodierleitfaden müssen vom Forscher festgelegt werden (Mayring 2015: 120). Dies ist wichtig zu erwähnen, da die Softwarelösungen komplexe Verbindungen zwischen den Daten erstellen und visualisieren können; die für das Forschungsproblem relevanten Relationen sowie die Interpretation der Ergebnisse sind aber weiterhin

82 NVivo 12 ist eine Software, die speziell für qualitativ orientierte Textanalysen entwickelt worden ist. QDA-Softwares werden häufig für qualitativ orientierte Textanalysen eingesetzt und können die komplexe Auswertung durch verschiedene Funktionen wirksam unterstützen (vgl. Mayring 2015: 116–122). Die Vorteile der computergestützten Verwendung sind u. a., dass die Software das Material übersichtlich darstellt und die Möglichkeit bietet, große und komplexe Datenmengen miteinander zu verbinden und zu bewältigen.

eine Aufgabe des Forschers (Bringer et al. 2004: 249). Des Weiteren wurden für einen Vergleich zwischen den Bewertungsperspektiven die qualitativen Daten quantifiziert, um analysieren zu können, in welchem Ausmaß die jeweiligen Aspekte in den schriftlichen Bewerterkommentaren vorkommen. Die Befunde der qualitativen Analysen zu von den Bewertenden in den schriftlichen Kommentaren beachteten Aspekten sind in Kap. 6 aufgeführt.

Qualitative Vergleiche der Bewerterurteile

Darüber hinaus wurden die Bewertungen im Hinblick auf die berücksichtigten Aspekte in den schriftlichen Kommentaren näher untersucht, um ergänzende Analysen zur zweiten und dritten Fragestellung vornehmen zu können. Hierbei wurden Vergleiche zwischen Urteilen zu Textproduktionen mit ähnlichen bzw. unterschiedlichen Ergebnissen anhand der qualitativen Inhaltsanalyse durchgeführt. Die Segmente der berücksichtigten Aspekte dieser Bewerterurteile wurden somit miteinander verglichen und in Verbindung gesetzt. Für die zweite Fragestellung hinsichtlich der Bewerterübereinstimmung zwischen den schwedischen Bewertenden wurden Bewerterurteile mit möglichst unterschiedlicher Benotung bzw. möglichst ähnlicher Benotung ausgewählt. Dabei wurde untersucht, inwiefern die Bewertenden in diesen Urteilen ähnliche oder unterschiedliche Aspekte berücksichtigen bzw. gewichten sowie inwieweit dies für die Benotung eine Bedeutung zu haben scheint. Hierzu wurde gleichzeitig geprüft, inwieweit sich Bewertende bei der Bewertung hauptsächlich auf Aspekte hinsichtlich einer Kategorie fokussieren oder auf Aspekte, die in vielen verschiedenen Kategorien einzuordnen sind.

Des Weiteren erfolgten für die dritte Fragestellung, die sich mit dem Verhältnis zwischen Bewertungen von schwedischen Bewertenden und GER-Bewertungen befasst, ähnliche Analysen. Hierfür wurden Bewerterurteile zu Textproduktionen auf *Tyska 5*, bei denen Unterschiede zwischen den Bewertungen der schwedischen Bewertenden im Hinblick auf das Erreichen eines Mindestanforderungsniveaus zu finden sind, und den GER-Bewertungen näher untersucht. Da Mindestanforderungen für ein erfülltes GER-Niveau B1 und eine ausreichende Note E auf *Tyska 5* in etwa äquivalent sein sollten, könnte eine qualitative Analyse dieser Bewerterurteile mehr Licht in dieses Verhältnis bringen. Der qualitativen Inhaltsanalyse der Bewerterkommentare, die in diesem Kapitel näher beschrieben wurde, liegen somit auch diese Analysen der Bewerterurteile zugrunde. Die Ergebnisse der qualitativen Vergleiche zwischen den Bewerterurteilen sind in Kap. 7.4 und 8.4 dargestellt.

5.3.2 Deskriptive Statistik und Korrelationsberechnungen

Grundsätzlich wurde im Hinblick auf die zweite Fragestellung bezüglich Bewerterübereinstimmung eine deskriptive Statistik für die Bewertungen auf *Tyska 3*, *Tyska 4* und *Tyska 5* berechnet. Hierbei wurden für die Ergebnisse der jeweiligen schwedischen Bewerterurteile Mittelwerte und Standardabweichungen pro Fremdsprachenstufe ermittelt. Diese Ergebnisse sind in Kap. 7.1 aufgeführt.

Um Aussagen über die dritte Forschungsfrage, d. h. die Beziehung der schwedischen Bewertungen zu einem erfüllten B1-Niveau des GER, treffen zu können, wurden in einem ersten Schritt die Ergebnisse der Schülertexte mithilfe deskriptiver Statistik quantitativ analysiert. Hierfür sind Extremwerte, Mittelwerte, Medianwerte und Standardabweichungen ermittelt worden. Der Anteil schwedischer Schülerleistungen, die die Anforderungen eines GER-Niveaus B1 erfüllt haben, wurde zudem pro Fremdsprachenstufe berechnet. Des Weiteren wurden die Ergebnisse in Bezug auf die Punktzahlen der jeweiligen GER-Bewertungen und die Benotung der schwedischen Bewertenden gemäß schwedischen Kriterien getrennt nach Fremdsprachenstufe aufgestellt.

Darüber hinaus wurden Korrelationen (Spearman's Rho) zwischen den jeweiligen Bewertungen berechnet, um das Verhältnis zwischen sämtlichen Bewertungen der beiden Bewertergruppen bzw. zwischen einzelnen Teilaspekten bei der Bewertung und den Gesamtbewertungen der schwedischen Bewertenden zu untersuchen. Die Spearman-Rangkorrelation wird am häufigsten verwendet, wenn Korrelationen zwischen Ratingwerten in Rangordnung bestimmt werden sollen (Wirtz & Caspar 2002: 133). Die Spearman-Rangkorrelation basiert auf Rangdaten und wird vor allem dann verwendet, wenn ordinale (wie bei den Bewertungen der schwedischen Bewertenden, die in der vorliegenden Studie den Leistungen die Noten F, E, D, C, B, und A gegeben haben) und nicht-normalverteilte Daten berechnet werden sollen.[83] Der Spearman-Koeffizient

83 Wenn viele Bewertungen denselben Noten zugeteilt worden sind, kann keine Rangordnung zwischen diesen Noten ermittelt werden, da identische Werte vorliegen. Eine Spearman-Rangkorrelation könnte demnach in diesem Fall problematisch sein, da die Bewertungen relativ häufig dieselben Noten bzw. Punktzahlen hinsichtlich einzelner Teilaspekten enthalten haben, sog. *ties* (vgl. Kendall & Dickinson Gibbons 1990: 40 ff.). Wenn dies der Fall ist, sollten die Signifikanzwerte von Spearman-Rangkorrelationen mit Vorsicht interpretiert werden. In der vorliegenden Studie sind aus diesem Grund zusätzlich die Korrelationsberechnungen zwischen schwedischen Bewertungen und GER-Bewertungen nach Kendalls Tau-b berechnet worden, um die Ergebnisse der Spearman-Rangkorrelationen durch ein ergänzendes Maß zu prüfen.

gibt ein Intervall zwischen r = -1 und r = 1 wieder, wobei r = 1 eine perfekte positive Korrelation zeigt und r = -1 eine perfekte negative Korrelation. Die quantitativen Untersuchungen der deskriptiven Statistik und der Korrelationskoeffizienten wurden mithilfe des Statistikprogramms SPSS berechnet. Die Ergebnisse zur Beziehung zwischen schwedischen Bewertungen und einem externen Referenzniveau sind in Kap. 8 zu finden.

5.3.3 Methoden zur Bestimmung der Bewerterübereinstimmung

Die zweite Forschungsfrage beschäftigt sich mit Differenzen der Bewertungen im Hinblick auf die Bewerterübereinstimmung. Um die Bewerterübereinstimmung untersuchen zu können, wurden in der vorliegenden Arbeit deskriptive Statistik sowie unterschiedliche Methoden zur Bestimmung der Bewerterübereinstimmung verwendet. Studien zur Beurteilerübereinstimmung werden häufig mittels Methoden der klassischen Testtheorie oder mittels IRT-Methoden vorgenommen. Zu einer grundlegenden Auffassung in der klassischen Testtheorie gehört die Variation von Bewerterurteilen. In der Diskussion des Forschungsfeldes zur Bewertung fremdsprachlicher Lernerproduktionen sind unterschiedliche Methoden, um die Beurteilerkonsistenz zu bewerten auseinandergesetzt worden. Zu den neueren Methoden gehören u. a. IRT-Analysen, insbesondere durch sog. *Multifacetten-Rasch-Modelle*. In Untersuchungen zur Bewerterübereinstimmung stellen Multifacetten-Rasch-Analysen ein adäquates Werkzeug dar, weil sie auch Facetten wie Bewerterstrenge, Aufgabenschwierigkeit und Fähigkeiten der Testteilnehmenden berücksichtigen (vgl. Eckes 2019).

Für Berechnungen der Übereinstimmung zwischen unterschiedlichen Bewertenden gibt es demnach eine Vielfalt von statistischen Methoden, die jeweils bestimmte Eigenschaften haben und deren Berechnungen demzufolge auch jeweils unterschiedliche Informationen ermitteln können. Zu beachten ist dabei, dass keine einzelne einheitliche oder beste Methode dieser Berechnungen vorhanden ist, sondern die Auswahl angemessener und geeigneter Methoden hängt von der jeweiligen Untersuchung ab. Dementsprechend ist es generell nützlich, verschiedene Typen von Berechnungen zu ermitteln, um ein breites Bild des Korpus zu erhalten (vgl. Wirtz & Caspar 2002: 23 ff.; Stemler

---

Kendalls Tau-b ist weniger empfindlich gegen Rangbindungen und damit für diese Berechnungen ein stabileres Maß (Wirtz & Caspar 2002: 137). Die Berechnungen nach Kendalls Tau-b haben jedoch im vorliegenden Fall ähnliche Ergebnisse wie die Rangkorrelationen nach Spearman gezeigt.

2004). Die vorliegende Arbeit folgt hier der Einteilung von Stemler (2004) sowie Stemler und Tsai (2008) in drei Methodenkategorien: Konsensmethoden, Konsistenzmethoden und Methoden zur Messwerteinschätzung (vgl. Kap. 3.3).

Konsensmethoden untersuchen in welchem Ausmaß Bewertende zu gleichen Urteilen über die Fähigkeiten der Lernenden kommen, wobei die Ermittlung der exakten *prozentualen Übereinstimmung (PÜ)* zu den einfachsten und meistverwendeten Methoden zählt (vgl. Stemler 2004; Jönsson & Svingby 2007). Die prozentuale Übereinstimmung wird häufig paarweise berechnet, um untersuchen zu können, wie jeder einzelne Bewertende mit jedem anderen Bewertenden übereinstimmt und damit nicht die Urteile eines einzelnen Bewertenden für den Gesamtwert zu viel Gewicht bekommen (vgl. Wirtz & Caspar 2002: 49). Für die exakte prozentuale Übereinstimmung gelten nach Stemler (2004) und Stemler und Tsai (2008) häufig Werte ab 70 % als zufriedenstellend, wenn anhand einer Bewertungsskala mit 5–7 Stufen bewertet wird, mit weniger Stufen sollte die prozentuale Übereinstimmung jedoch nicht unter 90 % liegen (vgl. Stemler 2004).

Außer der prozentualen Übereinstimmung werden zur Konsensschätzungen zwischen Bewerterpaaren häufig zufallskorrigierte Übereinstimmungsmaße ergänzend verwendet, insbesondere *Cohens Kappa* und *Cohens gewichtete Kappa* (Wirtz & Caspar 2002: 55 ff.).[84] Als Faustregel wird für eine gute Übereinstimmung oft ein Kappa-Grenzwert zwischen .60 und .75 angegeben, aber auch Werte zwischen .40 und .60 können akzeptabel sein (vgl. Landis & Koch 1977; Wirtz & Caspar 2002: 59 ff.; Stemler & Tsai 2008).

Während Cohens Kappa die Übereinstimmung zwischen zwei Bewertenden ermittelt, berücksichtigt Cohens gewichtete Kappa auch den Grad der Nicht-Übereinstimmung zwischen den Bewertenden. Mit anderen Worten bedeutet dies folgendes: „a greater ‚penalty' can be applied if the two categories chosen by the raters are farther apart" (vgl. Vanbelle 2016: 399) und dabei wird jeder Zelle ein Gewicht zugeordnet (vgl. Cohen 1968). Cohens gewichtete Kappa ist daher für Ermittlungen von ordinalen Werten gut geeignet.

84 Da die Werte sich auf einer Ordinalskala befinden, d. h. zwischen den hier gegebenen Schulnoten besteht eine Rangordnung, somit sind Berechnungen mit Cohens gewichtetem Kappa in diesem Fall zu empfehlen. Bei den Berechnungen zu Cohens gewichtetem Kappa werden größere Abweichungen bei der Einstufung der jeweiligen Bewertenden stärker ins Gewicht fallen als kleine Abweichungen. Um eine Indikation darüber geben zu können, inwiefern die beobachtete Übereinstimmung durch Zufall erklärt werden kann, sind in der vorliegenden Arbeit beide Kappa-Koeffizienten berechnet worden.

Die Nichtübereinstimmung kann entweder durch lineare Gewichtung oder quadratische Gewichtung berechnet werden. In der vorliegenden Studie wird die lineare gewichtete Kappa verwendet, da diese Berechnungen statistische Vorteile zeigen (vgl. Vanbelle 2016) und weniger von der Anzahl der verschiedenen Notenkategorien beeinflusst sind (vgl. Brenner & Kliebsch 2009). Der Grenzwert für Cohens gewichtete Kappa liegt wie bei Cohens ungewichtetem Kappa zwischen .60 und .70 für eine gute Übereinstimmung (vgl. Wirtz & Caspar 2002).

Konsistenzmethoden gehen der Frage nach, in welchem Ausmaß die Bewertungen in Relation oder Rangfolge zueinander stehen. Zu den Konsistenzmethoden gehören Berechnungen von *Spearman-Rangkorrelation* (vgl. hierzu auch Kap. 5.3.2) und *Kendalls Tau-b*, gebräuchliche Einheiten, um Korrelationen zwischen Bewertungen zu bestimmen (vgl. Wirtz & Caspar 2002: 133 ff.), aber auch das üblicherweise verwendete Konsistenzmaß *Cronbachs Alpha*, eines der am häufigsten verwendeten Methoden, um die interne Konsistenz von Bewertenden zu ermitteln (vgl. Stemler 2004). Da Korrelationsanalysen nur die interne Rangfolge miteinbeziehen, können diese Werte, auch wenn die Ergebnisse der Bewertungen nicht ganz genau übereinstimmen, hoch liegen. Die Konsistenzmaße sollten daher auch mit Konsensmethoden kombiniert werden. Bei einer stark positiven Korrelation liegen die Koeffizienten dem Wert +1 nahe und für eine stark negative Beziehung dem Wert -1. Ein Wert nahe 0 bedeutet, dass keine Korrelation vorhanden ist. Ein Ergebnis bei 1 oder in der Nähe von 1 weist somit darauf hin, dass die Rangwerte der Bewertungen in der Reihenfolge bei einer einzelnen Bewertenden mit den Rangwerten bei einem anderen Bewertenden in Verbindung stehen. Für die Konsistenzwerte gelten in der Regel Werte ab .7 als reliabel und ab .8 als gut (vgl. Barrett 2001; Stemler 2004; Stemler & Tsai 2008).

Ziel der dritten Kategorie, Methoden zur Messwerteinschätzung, ist es, alle verfügbare Informationen bei einer Bewertung zu sammeln und in Modelle zu inkorporieren, die es erlauben, die Interaktionen zwischen verschiedenen Bewertenden, Prüfungsteilnehmenden und Items (Aufgaben) zu untersuchen. Die meistverwendeten Methoden hierfür sind Faktoranalysen, Multifacetten-Rasch-Analysen oder die Verwendung von *Generalizability Theory* (vgl. Stemler 2004). Insbesondere das Multifacetten-Rasch-Modell wird relativ häufig im Bereich des Sprachtestens verwendet (vgl. Bachman et al. 1995; Eckes 2015; 2019), da dieses Modell es erlaubt, genaue Informationen über bestimmte Facetten zu gewinnen und ihren Einfluss auf die Bewertungen zu untersuchen.

Bei einer *Multifacetten-Rasch-Analyse* werden durch Rasch-Modelle unterschiedliche Parameter eingeschätzt und hierbei können z. B. die Fähigkeiten der

Prüfungsteilnehmenden, die Strenge- bzw. Milde-Tendenz der Bewertenden, Aufgabenschwierigkeit sowie der Schwierigkeitsgrad unterschiedlicher Kriterien untersucht werden (vgl. Stemler 2004). Diese Methoden werden häufig im Bereich Testkonstruktion und Testentwicklung eingesetzt. Bei einer Multifacetten-Rasch-Analyse sollten *Infit* bzw. *Outfit Mean-Square-Statistiken* (MnSq) ermittelt werden, damit untersucht werden kann, inwiefern diese Werte zum Raschmodell passen. Die Werte können somit über den Grad der Konsistenz der einzelnen Bewertenden informieren, indem sie ermitteln, inwiefern die Bewertungen einzelner Bewertender größere Variationen zeigen, als vom Modell erwartet wird, oder nicht. Sowohl Infit- als auch Outfit-Werte haben einen Erwartungswert von 1. Die Faustregel für die Interpretation von Mean-Square-Statistiken besagt, dass sowohl Infit- als auch Outfitwerte im Bereich von 0.5 bis 1.5 liegen sollten (vgl. Linacre 2002). Diese Richtwerte für akzeptable Grenzwerte können jedoch je nach Fragestellung variieren (vgl. Fan & Bond 2019).[85]

Die in den Analysen verwendeten Daten sind die Noten der Bewertungen von Textproduktionen schwedischer Gymnasialschülerinnen und Schüler, die von den schwedischen Bewertenden nach schwedischen Kriterien gegeben wurden. Jeder Schülertext wurde von der praktizierenden Lehrkraft und von zwei unabhängigen externen Bewertenden eingestuft, die alle als Lehrkräfte in schwedischen Schulen waren. Zu bemerken dabei ist, dass die Gruppe von Gymnasiallehrkräften hier als eine Einheit betrachtet wird, aber nicht desto weniger aus mehreren Individuen besteht: Die Gruppe der schwedischen Lehrkräfte ist somit hier eine Gruppenvariable, bestehend aus achtzehn unterschiedlichen Lehrkräften. Es darf hierbei auch nicht vergessen werden, dass einige Lehrkräfte nur einmal im untersuchten Datensatz vorkommen, während andere einen größeren Teil des Materials ausmachen.

Zur Bestimmung der Urteilgenauigkeit wurden in einem ersten Schritt jeweils drei Konsens- und Konsistenzmethoden verwendet, die im Bereich fremdsprachlicher Bewertungen häufig verwendet werden. Dabei wurden folgende Konsensmaße ermittelt: die prozentuale Übereinstimmung (PÜ), Cohens Kappa (κ) sowie Gewichtetes Kappa (κw). Für die prozentuale Übereinstimmung wurde der Anteil der Fälle berechnet, wo zwei Bewertenden dasselbe Urteil vergeben. Hierzu sind ergänzend Berechnungen der Konsistenzmaße zu Spearman's Rho, Kendalls Tau-b und Cronbachs Alpha vorgenommen worden. Um die Gruppe der Lehrkräfte mit den jeweiligen externen schwedischen

85 Auch engere Richtwerte hinsichtlich der Infit- bzw. Outfitwerte von 0.7–1.3 sind zu finden (vgl. Stemler & Tsai 2008; Fan & Bond 2019)

Bewertenden vergleichen zu können, wurden diese Berechnungen paarweise ermittelt (vgl. Eckes 2011). Die quantitativen Untersuchungen zur Bewerterübereinstimmung wurden mittels des Statistikprogramms SPSS durchgeführt.

Darüber hinaus wurden in einem zweiten Schritt die Bewertungen der schwedischen Bewertenden paarweise in Kreuztabellen einander gegenübergestellt, um Unterschiede bei den Bewertungen aufklären zu können und die Tendenz zur Strenge, Mitte bzw. Milde betrachten zu können (vgl. Eckes 2004; 2011). Der Grad der Beurteilerstrenge kann aber auch durch eine Multifacetten-Rasch-Analyse modelliert und berechnet werden. Für die Bewertungen der schwedischen Bewertenden wurde eine Multifacetten-Rasch-Analyse mit den Facetten „schriftliche Sprachfähigkeit der jeweiligen Prüfungsteilnehmenden" und „Beurteilerstrenge" vorgenommen. Die Multifacetten-Rasch-Analysen sind mit dem Computerprogramm MINIFAC (Version 3.58.0, Linacre 2005), einer freien Version der Software FACETS, durchgeführt worden.[86] Die Ergebnisse zur Übereinstimmung und Bewerterübereinstimmung zwischen der Gruppe der Lehrkräfte und den jeweiligen schwedischen Bewertenden sind in Kap. 7 dargestellt.

5.4 Begrenzungen der Methodik

Abschließend soll auf einige Grenzen der Methodik in der vorliegenden Arbeit hingewiesen werden. Eine Begrenzung der Studie stellt die *Stichprobe* dar. Es handelt sich zum einen darum, dass die Teilnahme der Schulen und der Lehrkräfte auf Freiwilligkeit beruhte. Bei der Auswahl der Schulen und der Lehrkräfte könnte es sich somit um eine sog. *self-selection-bias* handeln, da es keine Zufallsprobe ist. Dies könnte dazu geführt haben, dass vorrangig engagierte und erfahrene Lehrkräfte an der Studie teilnahmen, was die Repräsentativität und die Generalisierbarkeit der Studie in Frage stellen könnte. Eine Alternative für das Sampling wäre es, zufallsbasiert mit den Schulen Kontakt aufzunehmen. Allerdings könnte dies zur Folge haben, dass eventuell nicht genug Texte, insbesondere aus der Fremdsprachenstufe *Tyska 5*, erhoben werden würden und dass Gymnasialschulen, die weit auseinanderliegen, aus Zeitgründen nicht hätten besucht werden können. Auch wenn die Lehrkräfte der Studie nicht

86 Der Computerprogramm MINIFAC erfüllt dieselbe Funktion wie FACET, kann aber nur eine begrenzte Anzahl von Bewertungen berechnen (eine Höchstgrenze von 2000 Bewertungen). Die Bewertungen der vorliegenden Arbeit liegen aber unter dieser Grenze. MINIFAC kann unter folgender Internetadresse heruntergeladen werden: https://www.winsteps.com/minifac.htm

zufallsbasiert ausgewählt sind, ist darauf geachtet worden, dass sie unterschiedliche Schulformen repräsentieren und aus verschiedenen Regionen Schwedens kommen und dass sowohl weibliche als auch männliche Bewertende in der Studie vertreten sind.

Ebenfalls als Begrenzung hinsichtlich der Stichprobe können die teilnehmenden Probanden und die somit relativ begrenzte Anzahl von erhobenen Schülerleistungen betrachtet werden. Die Probanden dieser Arbeit waren Schülerinnen und Schüler auf *Tyska 3*, *Tyska 4* und *Tyska 5*, die sich alle in demselben schulischen Kontext befanden und die bei der Datenerhebung für die Studie freiwillig teilnahmen. Auch wenn die Schülerinnen und Schüler mit sehr wenigen Ausnahmen an der Studie teilgenommen haben, ist die Anzahl relativ begrenzt. Der Test ist jedoch unter authentischen Bedingungen erhoben worden und entsprach einer realistischen Testsituation. Dies trug dazu bei, dass die Probanden die Prüfung ernst genommen haben. Trotzdem haben eventuell nicht alle Probanden ihre gesamte schriftliche Kompetenz gezeigt. Aufgrund der kleinen Stichprobengröße sind auch die Analysen im Hinblick auf die Replizierbarkeit und die Generalisierbarkeit mit Vorsicht zu betrachten – bei einer höheren Anzahl von Texten hätte die Studie eventuell andere Ergebnisse gebracht. Da die Teilnahme der Schülerinnen und Schüler auf der Zusage der Schulen und der jeweiligen Lehrkraft beruhte, kann auch die Stichprobe der Schülerleistungen im Hinblick auf die Repräsentativität in Frage gestellt werden.

Eine Grenze der Studie entsteht auch durch Charakteristiken des zugrundeliegenden *Tests des schriftlichen Ausdrucks*. Der Prüfungsteil testet ausschließlich die schriftliche Kompetenz und überdies durch den niveauspezifischen Ansatz nur das Erfüllen bzw. Nicht-Erfüllen eines B1-Niveaus. Der Test besteht aus drei unterschiedlichen Aufgaben, die verschiedene Schreibkompetenzen prüfen und stammt von einem Sprachinstitut, das regelmäßige Qualitätskontrollen durchführt (vgl. Kap. 5.2), was insgesamt die Reliabilität des Tests stärkt. Um eine Generalisierung der Ergebnisse ermöglichen zu können, müssten jedoch weitere Tests und zusätzliche Aufgabenstellungen, die andere Teile der Sprachkompetenz und weitere Sprachniveaus berücksichtigen, verwendet werden. Im Rahmen der vorliegenden Studie war es nicht möglich, weitere Tests zu verwenden oder andere Teile der Sprachkompetenz bzw. zusätzliche Sprachniveaus zu prüfen.

Eine dritte Begrenzung der Studie ergibt sich in Bezug auf die *Bewertenden*. Zum einen bezieht sich diese Beschränkung auf die relativ begrenzte Anzahl der praktizierenden Lehrkräfte und zum anderen auf die begrenzte Anzahl der externen Bewertenden. Bisherige Untersuchungen aus einem schwedischen

Schulkontext haben eine unterschiedliche Anzahl von Bewertenden verwendet (vgl. Erickson 2009; Skolinspektionen 2010; Borger 2018; Dalberg 2019). Aus praktischen Gründen wäre es jedoch innerhalb der vorliegenden Untersuchung schwierig gewesen, zusätzliche Deutschlehrkräfte oder externe Bewertende hinzuzuziehen.

Ein wichtiger Punkt ist zudem, dass die 18 schwedischen Lehrkräfte als Gruppenvariable berücksichtigt werden. Es ist aber anzunehmen, dass es eine Variation innerhalb dieser Gruppe gibt. Aufgrund der relativen schmalen Materialbasis und des Versuchs, Tendenzen auf Systemebene zu finden, werden in dieser Arbeit keine Ergebnisse über den Lernstand von einzelnen Schulen dargestellt und die Ergebnisse werden zudem nicht einzelnen Lehrkräften zugeordnet.

Da es sich herausgestellt hat, dass Hintergrundfaktoren der Bewertenden Einfluss auf die Bewertungen haben können (vgl. Kap. 4), sollte z. B. die Lehrerfahrung der Bewertenden bei der Interpretation der Ergebnisse beachtet werden. Es ist zudem zu beachten, dass die Bewertenden in ihren Bewerterurteilen möglicherweise durch die Teilnahme an einer Forschungsstudie beeinflusst waren (z. B. Gustafsson & Erickson 2013). Auch sollte beachtet werden, dass die jeweiligen Bewertenden womöglich nicht alle Aspekte, die sie bei der Bewertung wahrgenommen haben, kommentieren, und dass sie außerdem andere Aspekte als diejenigen, die sie in ihren Kommentaren angegeben haben, berücksichtigen könnten (vgl. Lumley 2002). Dies sollte bei der Interpretation der Analyse Berücksichtigung finden.

Eine weitere Begrenzung der Studie ist die Tatsache, dass sie *kontextgebunden* ist. Die Studie ist generell von kontextuellen Faktoren der jeweiligen Schulen begrenzt. Hierzu gehört, dass ein Teil der Datenerhebung am Ende des Schuljahres stattfinden musste, um die Sprachkompetenz der Lernenden nach einer abgeschlossenen Stufe untersuchen zu können. Dies hat dazu geführt, dass Schulen und Lehrkräfte wegen Arbeitsbelastung oder anderer Aktivitäten kurzfristig abgesagt haben. Dies entspricht zwar oft den realen Bedingungen an Gymnasien am Ende des Schuljahres, sollte jedoch bei der Analyse beachtet werden.

Ebenfalls als Limitation im Hinblick auf den Kontext der Studie ist zu beachten, dass der Test als ein klassischer „Papier- und Bleistift-Test" angeboten wurde. Dies hat dazu geführt, dass die schwedischen Lehrkräfte und die externen Bewertenden die Textproduktionen in unterschiedlichen Formaten erhalten haben. Während die praktizierenden Lehrkräfte Textproduktionen auf Papier bewertet haben, um eine möglichst authentische Bewertungssituation herzustellen, standen den externen Bewertenden digitalisierte Texte zur

Verfügung, damit nicht schlechte Kopien oder Notizen der Lehrkräfte ihre Bewertung beeinflussen konnten (vgl. Gustafsson & Erickson 2013). Dies sollte ebenfalls Beachtung finden.

Darüber hinaus haben die jeweiligen Bewertergruppen, ihrem gewohnten Bewertungsverlauf folgend, teilweise unterschiedliche Voraussetzungen gehabt. Diese sind nicht nur auf die jeweiligen Anforderungen hinsichtlich eines B1-Niveaus oder der jeweiligen Kriterien der einzelnen Fremdsprachenstufen zurückzuführen, sondern haben auch mit den unterschiedlichen Bewertungsverfahren zu tun. Die GER-Bewertenden haben ein eher aufgabenspezifisches und analytisches Bewertungsraster verfolgt, während die schwedischen Bewertenden generell eher aufgabenübergreifende und holistische Kriterien verwendet haben. Auch dies hat wahrscheinlich einen Einfluss auf die Bewerterurteile der jeweiligen Bewertenden gehabt.

Abschließend können die bereits erwähnten *Analysemethoden* als eine weitere Einschränkung betrachtet werden. Vor allem können die Kategorienbildung der qualitativen Daten und die Objektivität der Kodierung in Frage gestellt werden. Die Diskussionen der Kategorienbildung und der Interkodierübereinstimmung zweier unabhängiger Forscher deuten jedoch zugleich auf eine ausreichende Objektivität des Kodierverfahrens hin und stärken somit die Validität und Reliabilität der vorliegenden Studie. Obgleich die Anwendung einer qualitativen Inhaltsanalyse zu einem erhöhten Verständnis für die Konstruktkonzeptualiserung der Bewertenden und mehr Information über die Inferenz der Bewertung führen kann, hat der Einsatz dieser Methode ebenso Kritik erhalten: Wenn Prüfer eine schriftliche Begründung ihrer jeweiligen Bewertungen abgeben, kann dennoch nie sichergestellt werden, welche Aspekte und Gedanken die Bewertung beeinflusst haben. Lumley (2005) mahnt aus diesem Grund zur Vorsicht bei der Interpretation von Bewerteraussagen: „The process of justifying the scores pushes the raters to select thoughts that are accessible for articulation" (S. 299). Des Weiteren ist laut Lumley auch eindeutig klar, dass Kommentare von Prüfern bei der Beurteilung nicht alle Aspekte des Bewertungsprozesses abdecken können: „they [*die Bewertenden*] could never verbalise more than a fraction of the thoughts that pass through their heads when rating" (S. 304). Auch wenn ein qualitativer Ansatz aus mehreren Gründen empfehlenswert sein kann, sollte dies ebenfalls berücksichtigt werden. Auch bei den statistischen Berechnungen sollte beachtet werden, dass statistische Methoden bestimmte Eigenschaften haben und somit auch unterschiedliche Informationen ermitteln können. Es war die Absicht, mittels der Auswahl für diese Studie geeigneter und unterschiedlicher Methoden ein breites Bild der untersuchten Bewertungen gewährleisten zu können.

6. Analyse des Fokus der Bewertenden

Mit Blick auf die in der Einleitung vorgestellten Fragestellungen wurden schriftliche Textproduktionen, die im Rahmen eines schriftlichen Tests auf B1-Niveau entstanden sind, erhoben. Diese wurden nach den schwedischen Bewertungsstandards für *Moderna språk* bzw. nach auf den GER-Standards basierend Kriterien bewertet. Nach der Beschreibung des methodischen Vorgehens (Kap. 5), um die in der Einleitung gestellten Fragestellungen beantworten zu können, werden hier und im folgenden Kapitel die Ergebnisse der Analysen dargelegt. Hierbei werden die Ergebnisse analog zur Reihenfolge der Fragestellungen dargestellt. Dieses Kapitel widmet sich der Konstruktkonzeptualisierung der Bewertenden, indem es die erste Fragestellung hinsichtlich des Evaluierungsprozesses aufgreift: *Welche Aspekte auf der Ebene der Texte sind in den jeweiligen Bewerterurteilen besonders relevant für die Beurteilung und wie unterschieden sich die Urteile zwischen einzelnen Bewertenden und Bewertergruppen bezogen auf: a) die eigene Lehrkraft, b) die externen schwedischen Bewertenden sowie c) die GER-Bewertenden?* Auch wenn im Rahmen der vorliegenden Studie Unterschiede bei der Bewertung zwischen einer praktizierenden Lehrkraft einerseits und externen Bewertenden andererseits nicht im Zentrum stehen und dies keineswegs erschöpfend untersucht werden kann, ist auf diverse mögliche Besonderheiten der Studie zu verweisen: Berücksichtigen schwedische Bewertende einerseits und GER-Bewertende andererseits in etwa die gleichen oder ähnliche Aspekte? Scheinen die jeweiligen Bewertenden dabei die beachteten Aspekte unterschiedlich zu gewichten oder nach denselben Maßstäben zu bewerten?

Im diesen Kapitel folgen die Ergebnisse der qualitativen Inhaltsanalyse der schriftlichen Kommentare, die von den Bewertenden der vorliegenden Studie als Begründungen für die Bewertungen der jeweiligen Schülerleistungen formuliert wurden. Diese Kommentare wurden nach einem Kodierverfahren qualitativ in unterschiedliche Kategorien eines Kodierungsschemas eingeordnet und analysiert. Die Befunde zu den beachteten Aspekten bei der Bewertung werden in der Analyse deskriptiv dargestellt, häufig aber aufgeteilt auf die beiden Bewertergruppen, d. h. die schwedischen Bewertenden bzw. die GER-Bewertenden. Im ersten Teil wird auf die Verteilung derjenigen Aspekte eingegangen, die in den Urteilen der jeweiligen Bewertenden zum Vorschein kommen. Hierbei werden sowohl Parallelen als auch Unterschiede dieser beachteten Bewerteraspekte zwischen den in der vorliegenden Arbeit betrachteten Bewertergruppen dargelegt (Kap. 6.1). Anschließend wird die Verteilung negativer, gemischter und positiver Kommentare pro Kategorie dargestellt

(Kap. 6.2). Danach wird eine vertiefte Analyse im Hinblick auf die jeweiligen Kategorien der Inhaltsanalyse vorgenommen, wobei Beispiele der Bewerterkommentare zur Illustration der verschiedenen Bewerteraspekte präsentiert werden (Kap. 6.3). Ein Fazit fasst die wichtigsten Ergebnisse zusammen und beschließt das Kapitel (Kap. 6.4).

6.1 Verteilung der Bewerterkommentare pro Kategorie

Im Folgenden wird die Frage bearbeitet, inwiefern die Bewertenden eine Variabilität im Hinblick auf bedeutsame Aspekte bei der Bewertung aufweisen. Untersucht wird auch, inwiefern Bewertende in ihren Urteilen gewissen Aspekten mehr Gewicht verleihen und es dabei Unterschiede in Bezug auf verschiedene Bewertergruppen gibt. Tab. 15 zeigt eine quantitative Zusammenfassung der qualitativen Inhaltsanalyse der schriftlichen Bewerterkommentare. Diese sind nach den Hauptkategorien kodiert (vgl. Kap. 5.3 für die Ausarbeitung und Analyseprinzipien des Kodierschemas) und auf die beiden Bewertergruppen, d. h. die schwedischen Bewertenden (die Gruppe der Lehrkräfte und die zwei externen Bewertenden) bzw. die GER-Bewertenden, verteilt:[87]

Tab. 15: *Gesamtergebnis der beachteten Aspekte bei der Bewertung schriftlicher Kompetenz, Gesamtanzahl pro Kategorie und in Prozent angegeben (N = 300)*

Beachtete Aspekte	*schwed. Bewertende*		*GER-Bewertende*		***Gesamt***	
	N	**%**	**N**	**%**	**N**	**%**
Angemessenheit	156	13,9	293	28,7	449	21,0
Aufgabenerfüllung	147	13,1	201	19,7	348	16,3
formale Strukturen	198	17,7	204	20,0	402	18,8
Gesamteindruck	111	9,9	0	0,0	111	5,2
kommunikative Strategien	15	1,3	0	0,0	15	0,7
pauschale Beurt. – Sprache	47	13,1	1	0,1	148	6,9
Sonstiges	19	1,7	1	0,1	20	0,9
Textfluss	37	3,3	0	0,0	37	1,7
Verständlichkeit	122	10,9	112	11,0	234	10,9
Wortschatz	167	14,9	208	20,4	375	17,5
Gesamt	**1 119**	**100**	**1 020**	**100**	**2 139**	**100**

87 Hierbei ist zu beachten, dass die Werte der schwedischen Bewertenden (der Gruppe der Lehrkräfte und der zwei externen Bewertenden) sich auf 180 Bewerterurteile beziehen, wohingegen die Werte der zwei GER-Bewertenden auf 120 Bewerterurteile bezogen sind. Aus diesem Grund werden die Ergebnisse der beiden Bewertergruppen hier getrennt behandelt.

Aus der Tabelle ist zu erkennen, dass die schwedischen Bewertenden ein breiteres Spektrum von Aspekten beachten, wobei einige Kategorien jedoch im Vergleich zu den anderen überwiegen. In ihren Begründungen machen Aspekte der *formalen Strukturen* (17,7 %) den größten Anteil aus. Darüber hinaus sind aber auch Aspekte zum *Wortschatz* (14,9 %), zur *Angemessenheit* (13,9 %), zur *Aufgabenerfüllung* (13,1 %) sowie zu einer *pauschalen Beurteilung der Sprache* (13,1 %) zu verzeichnen, die in etwa in gleichem Ausmaß beachtet werden. Zudem kommen in den Bewerterkommentaren der schwedischen Bewertenden Aspekte, die den Kategorien *Verständlichkeit* (10,9 %) und *Gesamteindruck* (9,9 %) zugeordnet werden können, zum Ausdruck. Geringer ist der Gesamtanteil von Aspekten, die zu den Kategorien *Textfluss* (3,3 %), *kommunikative Strategien* (1,3 %) sowie *Sonstiges* (1,7 %) gehören und die insgesamt nur wenige Prozente ausmachen.

Die berücksichtigten Aspekte in den Kommentaren der GER-Bewertenden sind hauptsächlich auf fünf der Kategorien verteilt. Der Aspekt der *Angemessenheit* macht in den Bewerterkommentaren der GER-Bewertenden den größten Anteil aus, fast ein Drittel der beachteten Aspekte gehören zu dieser Kategorie (28,7 %). Des Weiteren bestehen jeweils ein Fünftel der Kommentare aus Aspekten, die den Hauptkategorien *Wortschatz, formale Strukturen* und *Aufgabenerfüllung* (jeweils etwa 20 %) zugeordnet werden können. Zu den meistbeachteten Aspekten der GER-Bewertenden gehören auch Kommentare, die auf die *Verständlichkeit* (11 %) in den Schülertexten zurückzuführen sind. Im Gegensatz zu den schwedischen Bewertenden geben die GER-Bewertenden sehr selten eine pauschale Bewertung der Sprache ab und kommentieren gar nicht den Gesamteindruck, kommunikative Strategien oder den Textfluss in den Schülerleistungen.

Aus der Tabelle wird zudem deutlich, dass die Gesamtanzahl der Kommentare der schwedischen Bewertenden nur ganz wenig höher als die der GER-Bewertenden ist, 1119 gegen 1020. Dies bedeutet, dass schwedischen Bewertenden pro Person weniger Aspekte beachten im Vergleich zu den GER-Bewertenden. Eines der Ergebnisse der vorliegenden Studie ist, dass die GER-Bewertenden quantifizierbar mehr Aspekte der sprachlichen Kompetenz pro Schülertext kommentieren als die schwedischen Bewertenden.

Zur Illustration wird die Distribution der Aspekte in den Kommentaren beider *Bewertergruppen* (d. h. zwischen den schwedischen Bewertenden und den GER-Bewertenden) in den einzelnen Hauptkategorien in Abb. 8 einander vergleichend gegenübergestellt:

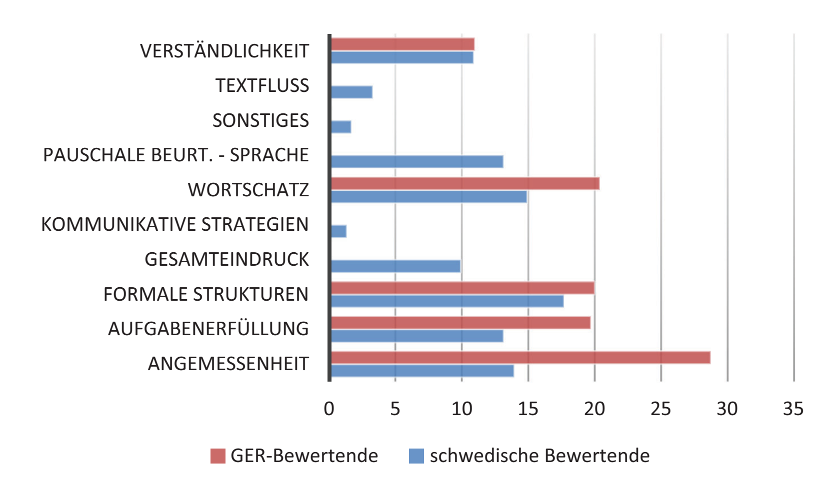


Abb. 8: Verteilung der Bewerterkommentare auf die Hauptkategorien, schwedische Bewertende ($N = 180$) bzw. GER-Bewertende ($N = 120$) im Vergleich, in Prozent angegeben

Bei der Verteilung der beachteten Aspekte lassen sich Ähnlichkeiten, aber auch deutliche Diskrepanzen zwischen den beiden Bewertergruppen erkennen. Die Kategorie *Angemessenheit* ist der meistbeachtete Aspekt in den Bewerterurteilen der GER-Bewertenden (in der Abbildung rot dargestellt), gefolgt von den Kategorien *Wortschatz*, *formalen Strukturen* und *Aufgabenerfüllung*. An erster Stelle bei den schwedischen Bewertenden (in der Abbildung blau dargestellt) steht dagegen die Kategorie *formale Strukturen*, ein Aspekt, der eher indirekt in den schwedischen Bildungsstandards vorkommt. Des Weiteren widmen die GER-Bewertenden ihre Aufmerksamkeit in höherem Maße den Aspekten *Wortschatz*, *Angemessenheit* und *Aufgabenerfüllung*. Selbst wenn häufig die gleichen Aspekte von den Bewertenden beachtet werden, finden sich auch deutliche Unterschiede: So werden Kommentare der GER-Bewertenden hauptsächlich fünf Kategorien zugeordnet, während die Kommentare der schwedischen Bewertenden eine deutlich breitere Verteilung zeigen. Eine Ausnahme ist der Aspekt *Verständlichkeit*, der von sowohl den schwedischen als auch den GER-Bewertenden in etwa gleichem Umfang berücksichtigt wird und rund 11 % der Kommentare in beiden Bewertergruppen ausmacht.

Weitere Unterschiede zwischen den einzelnen Gruppen können bei gewissen Aspekten beobachtet werden. Hierzu gehören Kommentare zu Kategorien, die von den GER-Bewertenden kaum berücksichtigt werden, z. B. die *pauschale Beurteilung – Sprache* sowie der *Gesamteindruck*. Diese Aspekte werden von den schwedischen Bewertenden vergleichsweise häufig beachtet. Der Vergleich

der beiden Bewertergruppen zeigt somit, dass schwedische Bewertende in höherem Ausmaß generelle Aussagen über Schülerleistungen treffen, teils über einen globalen Gesamteindruck der Lernproduktionen und teils über die Sprache in den Schülertexten. Darüber hinaus beachten schwedische Bewertende in ihren Bewertungen in gewissem Ausmaß auch Aspekte wie *Textfluss* oder *kommunikative Strategien*, Kategorien, die von den GER-Bewertenden überhaupt nicht aufgegriffen werden.

Die Ergebnisse zur Distribution der Kommentare weisen zusammenfassend darauf hin, dass die Herangehensweise der schwedischen Bewertenden mit sich bringt, dass ein breiteres Spektrum von Aspekten berücksichtigt wird, wobei zu den Schülertexten ein Globalurteil, häufig ohne ein vorgelegtes Muster, abgegeben wird. Die GER-Bewertenden verwenden dagegen ein Bewertungsraster, das vier Bewertungsdimensionen enthält, was ein anderes Vorgehen verlangt. Dies scheint sich in den von den GER-Bewertenden berücksichtigten Aspekten wiederzuspiegeln.

6.2 Verteilung positiver, gemischter bzw. negativer Bewerterkommentare

Um weitere Tendenzen erkennen zu können und zwischen eventuellen Stärken und Schwächen in den Schülerleistungen zu unterscheiden, wurden die Segmente in positive, gemischte und negative Kommentare aufgeteilt. Die Ergebnisse dieser Analyse pro Kategorie lassen sich für sämtliche Bewertende Tab. 16 entnehmen:

Tab. 16: *Verteilung der positiven, gemischten bzw. negativen Segmente pro Hauptkategorie (Anzahl der Segmente und in Prozent angegeben) (N = 300)*

Beachtete Aspekte	*positiv*		*gemischt*		*negativ*	
	N	**%**	**N**	**%**	**N**	**%**
Angemessenheit	121	27,0	134	29,8	194	43,2
Aufgabenerfüllung	78	22,4	109	31,3	161	46,3
formale Strukturen	125	31,1	193	48,0	84	20,9
Gesamteindruck	20	18,0	24	21,6	67	60,4
kommunikative Strategien	2	13,3	1	6,7	12	80,0
pauschale Beurt. – Sprache	71	48,0	56	37,8	21	14,2
Sonstiges	7	35,0	2	10,0	11	55,0
Textfluss	2	5,4	6	16,2	29	78,4
Verständlichkeit	67	28,6	88	37,6	79	33,8
Wortschatz	102	27,2	124	33,1	149	39,7
Gesamt	**595**	**27,8**	**737**	**34,5**	**807**	**37,7**

In Tab. 16 ergeben sich große Unterschiede zwischen den jeweiligen Aspekten im Hinblick darauf, inwiefern sie in positiven, gemischten oder negativen Worten erfasst werden. Die Mehrheit der Segmente sind insgesamt negativ ausgewertete Kommentare (etwa 38 %). Dies könnte auf eine leichte Tendenz, Defizite in den Textproduktionen zu kommentieren, hindeuten. Nichtdestoweniger ist eine nicht zu vernachlässigende Anzahl der Segmente, insgesamt etwa zwei Drittel, entweder als positiv (etwa 28 %) oder gemischt (etwa 35 %), d. h. weder positiv noch negativ bzw. zweideutig kodiert, was aber auch in eine andere Richtung deutet.

Ein überwiegender Anteil der negativen Kommentare findet sich bei Aspekten, die in die Kategorien *Gesamteindruck, Angemessenheit*, *Aufgabenerfüllung, Wortschatz* und *Verständlichkeit* einzuordnen sind. Auffallend ist insbesondere die große Zahl von negativ kodierten Segmenten, die zu den Aspekten *Angemessenheit* (N = 194), *Aufgabenerfüllung* (N = 161) und *Wortschatz* (N = 149) zurückzuführen sind. Zu den negativen Kommentaren gehören u. a. Aussagen über Mängel hinsichtlich der formellen Anrede oder der Textsorte, das Fehlen einer Teilaufgabe sowie Schwierigkeiten der Wortschatzbeherrschung. Auch wenn die negativen Einschätzungen bei einigen der übrigen Kategorien überwiegen, handelt es sich hier um eine geringere Anzahl von Kommentaren, u. a. bei den Kategorien *Sonstiges*, *kommunikative Strategien* und *Textfluss*, was bei der Analyse betrachtet werden muss.

Unter den Segmenten, die als gemischt kodiert sind, stellen Aspekte, die *formalen Strukturen* zugeordnet werden können, den höchsten Anteil dar: etwa die Hälfte sämtlicher Segmente in dieser Kategorie sind als gemischt kodiert. Einen großen Anteil gemischt kodierter Segmente haben auch die Kategorien *pauschale Beurteilung der Sprache* und *Verständlichkeit.* Ein überwiegender Anteil der Kommentare, die einer *pauschale Beurteilung der Sprache* zugeordnet werden können, ist aber positiv. Eine generelle Aussage über die Sprache scheint daher häufiger in positiven Worten oder sowohl Schwächen als auch Stärken betreffend beschrieben zu werden. Auch Aspekte, die zu *formalen Strukturen* zurückzuführen sind, haben insgesamt einen relativ großen Anteil positiv kodierter Aspekte. Inwiefern sich die Bewertergruppen im Hinblick darauf unterschieden, in welchem Ausmaß sie positive, negative oder gemischte Kommentare geben und zu welchen Aspekten, wird im nächsten Abschnitt näher beschrieben.

6.3 Analyse der Bewerterkommentare pro Kategorie

Der folgende Abschnitt enthält eine vertiefende quantitative und qualitative Analyse bezüglich der in den Bewerterkommentaren berücksichtigten Aspekte. Diese sind den jeweiligen Hauptkategorien, gegebenenfalls inklusive ihrer

Subkategorien, zugeordnet. Hierbei wird die Verteilung der positiven, gemischten und negativen Kommentare durch Abbildungen und Tabellen pro Kategorie/Subkategorie ersichtlich. Anschließend werden illustrierende Beispiele der schriftlichen Bewerterkommentare gegeben, um die Facetten der jeweiligen Bewertungsdimensionen zu beschreiben und zu veranschaulichen. Auf kennzeichnende und deutliche Unterschiede zwischen einzelnen Bewertenden oder Bewertergruppen wird explizit eingegangen.

Weitgehend werden die Hauptkategorien einzeln dargestellt, damit die Kommentare der jeweiligen Aspekte einfacher zu vergleichen sind. Da die Subkategorien der *formalen Strukturen* und des *Wortschatzes* viele Gemeinsamkeiten aufweisen, sind sie hier gemeinsam unter *Aspekten der linguistischen Kompetenz* aufgeführt und beleuchtet. Hinzu kommt dabei auch die *pauschale Beurteilung der Sprache*. Aspekte, die in diese Kategorie einzuordnen sind, finden sich hauptsächlich in den Kommentaren der schwedischen Bewertenden, während die GER-Bewertenden häufig Aspekte zum *Wortschatz* und zu *formalen Strukturen* getrennt kommentieren. Darüber hinaus sind Aspekte der vier Hauptkategorien (*Gesamteindruck*, *Textfluss*, *kommunikativen Strategien* und *Sonstiges*), die von den GER-Bewertenden kaum berücksichtigt werden, in einem Kapitel zusammengeführt.

Zunächst werden dementsprechend die Befunde zum Bewertungsprozess im Hinblick auf *Aspekte der linguistischen Kompetenz* (Kap. 6.3.1), Aspekte zur *Verständlichkeit* (Kap. 6.3.2), Aspekte zur *Aufgabenerfüllung* (Kap. 6.3.3), Aspekte zur *Angemessenheit* (Kap. 6.3.4) sowie Aspekte zum *Gesamteindruck*, zum *Textfluss*, zu *kommunikativen Strategien* sowie zu *Sonstigem* (Kap. 6.3.5) dargelegt und vertieft.

6.3.1 Aspekte der linguistischen Kompetenz

Die Kommentare zu ***Aspekten der linguistischen Kompetenz*** sind in die drei Hauptkategorien *formale Strukturen*, *Wortschatz* sowie *pauschale Beurteilung – Sprache* eingeteilt. Für die Kategorien *formale Strukturen* und *Wortschatz* ergibt sich eine weitere Einteilung in Subkategorien zur *Korrektheit und Präzision* sowie zur Bandbreite (als *Spektrum* bzw. *Differenziertheit* realisiert). Die Hauptkategorien *formale Strukturen* und *Wortschatz* enthalten zudem jeweils eine domänenspezifische Unterkategorie: *Orthographie* (*formale Strukturen*) und *idiomatische Ausdrücke* (*Wortschatz*). Tab. 17 und 18 zeigen die Verteilung dieser Aspekte auf Haupt- und Subkategorien, wobei die Verteilung auf positive, gemischte oder negative Segmente pro Kategorie dargestellt ist. Die Ergebnisse sind nach den beiden Bewertergruppen, den schwedischen Bewertenden bzw. den GER-Bewertenden, aufgeteilt:

Tab. 17: *Verteilung der Kommentare der schwedischen Bewertenden (N = 180) auf Aspekte der linguistischen Kompetenz (Anzahl der Segmente)*

Aspekte der linguistischen Kompetenz	*negativ*	*gemischt*	*positiv*	***Gesamt***
– Korrektheit – Präzision	94	20	16	130
– Orthographie	46	0	4	50
– Spektrum	1	1	16	18
formale Strukturen – gesamt	**141**	**21**	**36**	**198**
– Differenziertheit	9	6	34	49
– idiomatische Ausdrücke	16	6	21	43
– Korrektheit – Präzision	63	4	8	75
Wortschatz – gesamt	**88**	**16**	**63**	**167**
pauschale Beurt. – Sprache	**55**	**71**	**21**	**147**
Gesamt	**284**	**108**	**120**	**512**

Tab. 18: *Verteilung der Kommentare der GER-Bewertenden (N = 120) auf Aspekte der linguistischen Kompetenz (Anzahl der Segmente)*

Aspekte der linguistischen Kompetenz	*negativ*	*gemischt*	*positiv*	***Gesamt***
– Korrektheit – Präzision	15	60	35	110
– Orthographie	25	5	0	30
– Spektrum	12	39	13	64
formale Strukturen – gesamt	**52**	**104**	**48**	**204**
– Differenziertheit	9	36	53	98
– idiomatische Ausdrücke	1	0	1	2
– Korrektheit – Präzision	26	50	32	108
Wortschatz – gesamt	**36**	**86**	**86**	**208**
pauschale Beurt. – Sprache	**1**	**0**	**0**	**1**
Gesamt	**89**	**190**	**134**	**413**

Wie aus den Tabellen ersichtlich überwiegen generell Segmente, die auf ***formale Strukturen*** und ***Wortschatz*** zurückzuführen sind. Darunter gilt für sowohl die schwedischen als auch die GER-Bewertenden, dass ***Korrektheit und Präzision*** die meistbeachtete Subkategorie bildet. Auffallend hierbei ist aber die verhältnismäßig große Diskrepanz hinsichtlich der Verteilung der Kommentare, die als *Korrektheit/Präzision* einzuordnen sind im Vergleich zu Kommentaren, die auf die Bandbreite (***Spektrum/Differenziertheit***) dieser beiden Kategorien verweisen. Das Spektrum der *formalen Strukturen* wird zwar relativ häufig von den GER-Bewertenden kommentiert, betrifft jedoch nur die Hälfte der Segmente zur *Korrektheit und Präzision*. Ebenfalls wird die Anzahl von

Segmenten hinsichtlich der Subkategorie *Spektrum* (*formale Strukturen*) von den schwedischen Bewertenden weniger beachtet. Die entsprechende Subkategorie *Differenziertheit* (*Wortschatz*) unterscheidet sich aber in geringerem Ausmaß von Wortbeherrschung und Korrektheit. Die Bandbreite und Variation innerhalb der *formalen Strukturen* scheinen somit für die Bewertenden im Vergleich zur Differenziertheit des *Wortschatzes* bei der Bewertung geringere Aufmerksamkeit zu erhalten.

Des Weiteren wird die domänenspezifische Subkategorie ***Orthographie*** (*formale Strukturen*) von beiden Bewertergruppen kommentiert, scheint aber von den schwedischen Bewertenden in etwas größerem Ausmaß Aufmerksamkeit zu erhalten. Dies liegt offensichtlich daran, dass einige der schwedischen Bewertenden im Material in den Textproduktionen häufiger die orthographischen Abweichungen korrigieren. Im gleichen Sinne wird die zweite domänenspezifische Subkategorie ***idiomatische Ausdrücke*** (*Wortschatz*) in höherem Ausmaß von den schwedischen Bewertenden als von den GER-Bewertenden kommentiert. Unter den schwedischen Bewertenden wird dieser Aspekt fast im gleichen Ausmaß wie die lexikalische Differenziertheit kommentiert, während die GER-Bewertenden sehr selten Kommentare über Phrasen oder Ausdrücke in den Schülerleistungen abgeben. Die größte Diskrepanz zwischen den Bewertergruppen ergibt sich dennoch bei der Kategorie ***pauschale Beurteilung – Sprache***. Wie aus den Tabellen oben ersichtlich, werden Kommentare dieser Art hauptsächlich von den schwedischen Bewertenden zum Ausdruck gebracht. Die GER-Bewertenden dagegen verzichten in ihren Urteilen mit nur einer Ausnahme auf generelle Aussagen über die Sprache in den Textproduktionen.

Weitere Unterschiede zwischen den Bewertergruppen betreffen die *Einteilung in positive, negative und gemischte Segmente*. Sehr auffallend ist hierbei, dass schwedische Bewertende insgesamt in bedeutend höherem Ausmaß gerade Aspekte der linguistischen Kompetenz in den Textproduktionen negativ einschätzen: insgesamt ist über die Hälfte der Segmente in den schwedischen Urteilen zu diesen Aspekten negativ bewertet (vgl. Tab. 17). Dies gilt vor allem bezüglich grammatischer Mängel und der Rechtschreibung (*formale Strukturen*) sowie der Wortwahl und Wortschatzbeherrschung (*Wortschatz*). Dies zeigt sich auch wenn die schwedischen Bewertenden eine *pauschale Bewertung der Sprache* geben.

Ein teilweise anderes Bild wird vermittelt, wenn Aspekte, sich auf die Bandbreite der beiden Kategorien oder auf Phrasen beziehen. In den Bewerterurteilen wird ersichtlich, dass schwedische Bewertende bezüglich der Subkategorie *formale Strukturen – Spektrum* fast ausschließlich positiv bewerten. Des Weiteren überwiegen auch bei den Subkategorien *Differenziertheit* (*Wortschatz*) und

idiomatische Ausdrücke (*Wortschatz*) positive Einschätzungen. Insgesamt fällt bei den schwedischen Bewertenden auf, dass die negativ bewerteten Aspekte auf sprachliche Korrekturen der Rechtschreibung bzw. auf der Wort- und Satzebene zurückzuführen sind, während positive Kommentare im Hinblick auf die Bandbreite im lexikalischen oder grammatischen Bereich überwiegen.

Unter den GER-Bewertenden werden Aspekte der linguistischen Kompetenz bei der Bewertung generell höher eingeschätzt, bei ihnen überwiegen hier die positiven und gemischten Einschätzungen über die negativen. Die Mehrheit der Kommentare zu diesen Aspekten sind aber gemischter Art, was darauf hindeutet, dass die GER-Bewertenden Aspekte der linguistischen Kompetenz in den Lernproduktionen häufig im mittleren Bereich des Bewertungsrasters (vgl. Anhang 12) ansiedeln. Ihre Kommentare geben hierbei an, dass Fehlgriffe in diesen Bereichen zwar zu finden sind, aber dass sie nur teilweise das Verständnis beeinflussen. Wie bei den schwedischen Bewertenden überwiegen die positiven Kommentare hinsichtlich der *lexikalischen Differenziertheit* und die negativen bezüglich der *Orthographie*.

Des Weiteren lassen sich weitere Unterschiede zwischen der Gruppe der schwedischen Lehrkräfte und den externen schwedischen Bewertenden erkennen. Auffallend ist vor allem die Tendenz, dass die schwedischen Deutschlehrkräfte[88] sprachliche Korrekturen im Vergleich zu den beiden externen Bewertenden häufiger kommentieren. Die Verteilung der negativen Kommentare im Hinblick auf sprachliche Korrekturen, d. h. zur *Korrektheit/Präzision* (*formale Strukturen* und *Wortschatz*) sowie der *Orthographie* (*formale Strukturen*), ist in Tab. 19 wiedergeben:

Aus Tab. 19 ergibt sich, dass die schwedischen Deutschlehrkräfte eine gewisse Tendenz haben, sprachliche Korrekturen in höherem Ausmaß als die externen schwedischen Bewertenden zu beachten. In diesem Zusammenhang scheinen somit die Deutschlehrkräfte einen etwas höheren Fokus auf grammatische Mängel, orthographische Fehlgriffe sowie Wortschatzbeherrschung zu haben. Die zwei schwedischen externen Bewertenden weisen bei den Kommentaren zur Orthographieeine ähnliche Verteilung auf, unterscheiden sich aber untereinander teilweise im Hinblick auf die Präzision der formalen Strukturen und des Wortschatzes. Verglichen mit den externen schwedischen Bewertenden kann somit eine Diskrepanz bezüglich sprachlichen Korrekturen wahrgenommen werden. Die GER-Bewertenden weisen diesbezüglich keine großen Differenzen auf.

88 Wie in Methodikkapitel anbemerkt, ist die Gruppe der Lehrkräfte hier als eine Gruppenvariable dargestellt, besteht allerdings aus mehreren Individuen ($N = 18$).

Tab. 19: *Verteilung der negativen Bewerterkommentare der Gruppe der schwedischen Bewertenden (N = 180) auf sprachliche Korrekturen (Anzahl der Segmente)*

Sprachliche Korrekturen	*Gruppe der Lehrkräfte*	*ext. schwed. Bewert. 1*	*ext. schwed. Bewert. 2*	***Gesamt***
Korrektheit – Präzision (formale Strukt.)	43	31	20	**94**
Orthographie (formale Strukt.)	38	5	3	**46**
Korrektheit – Präzision (Wortschatz)	37	10	16	**63**
Gesamt	**118**	**46**	**39**	**203**

Formale Strukturen: Korrektheit/Präzision

In der qualitativen Analyse der Bewerterkommentare können sowohl Gemeinsamkeiten als auch Unterschiede bezüglich der Aspekte der linguistischen Kompetenz bemerkt werden. In den Bewerterurteilen zur Subkategorie ***Korrektheit/Präzision*** (*formale Strukturen*) kommentierten die Bewertenden auf der einen Seite einen globalen Eindruck grammatischer Fehlgriffe, siehe Beispiele 6.1 und 6.2:

(6.1) med en del grammatikfel[89] (Hjbt5-3-E, Lehrkraft)[90]

(6.2) Strukturen: mehrere Fehlgriffe beeinträchtigen das Verständnis erheblich. (Hjbt5-3, GER-Bewert. 2)

Schwedische Bewertende geben dabei oft eine pauschale Bewertung zur grammatischen Korrektheit in den Texten (vgl. Beispiel 6.1), während die GER-Bewertenden häufiger auf Fehlgriffe und deren Einfluss auf das Verständnis verweisen (vgl. Beispiel 6.2). Die GER-Bewertenden setzen generell, dem Bewertungsraster folgend, häufig die sprachliche Korrektheit in Verbindung mit der Verständlichkeit und geben Kommentare darüber ab, inwiefern formale Fehlgriffe das Verständnis der Leistung beeinträchtigen.

Die Bewertenden kommentieren auch spezifische grammatische Phänomene in den Schülertexten. Sie verweisen hierbei häufig auf grammatische Phänomene, wie die Bewältigung von Satzstellung und Verbformen. In gewissem Ausmaß kommentieren sie auch die Beherrschung von Genus, Kasus und Adjektivbeugung. Für diese Phänomene sind hier einige Beispiele aus den Urteilen aufgeführt (vgl. Beispiele 6.3–6.6):

89 „Mit einigen Grammatikfehlern". (*Hier und im Folgenden eigene Übersetzung, M.H.R.*).

90 Alle Zitate der Bewertenden werden im Folgenden originalgetreu wiedergegeben.

(6.3) Schwierigkeiten mit der Satzstellung (Sätze mit modalen Hilfsverben und Nebensätzen), bzw. Nichtbeherrschen der Passivkonstruktion (will nicht publiziert werden) deren Kenntnis auf B1 Niveau erwartet werden kann." (Cswu1-3, GER-Bewert. 1).
(6.4) Dock många språkliga missar gällande enklare verbformer som borde sitta.[91] (Hjbt5-3-E, ext. schwed. Bewert. 2).
(6.5) Fixar bisatsordföljd två gånger men dras ändå med vissa enklare verbfel.[92] (Ghhs4-3-D, ext. schwed. Bewert. 1).
(6.6) Ordföljden något osäker (t.ex. i bisatser, verbböjning)[93] (Geks8-3-E, Lehrkraft)

Wie aus den Beispielen 6.3 und 6.4 ersichtlich, kommentieren sowohl die schwedischen Bewertenden als auch die GER-Bewertenden gelegentlich Schwierigkeiten im Hinblick auf grammatische Phänomene, von denen sie meinen, dass Lernende auf einem gewissen Niveau sie bewältigen können sollten. Dennoch sind deutliche Unterschiede zwischen den beiden Bewertergruppen zu erkennen: Schwedische Bewertende kommentieren häufiger die Korrektheit einzelner grammatischer Phänomene. Die schwedischen praktizierenden Lehrkräfte schienen auch, wie bereits oben erwähnt, im Vergleich zu den beiden externen schwedischen Bewertenden in höherem Grad ihre Aufmerksamkeit auf sprachliche Korrekturen zu richten. Sie sind oft detaillierter, wobei sie etwas häufiger in Fehlertypen kategorisieren und hierbei sowohl morphologische als auch syntaktische Phänomene berücksichtigen (vgl. Beispiel 6.6). Die GER-Bewertenden dagegen verwenden oft die Formulierung im Bewertungsraster (vgl. Beispiel 6.2), kommentieren aber auch gelegentlich grammatische Schwierigkeiten (vgl. Beispiel 6.3).

Formale Strukturen: Spektrum

In die Subkategorie ***Spektrum*** (*formale Strukturen*) sind im Vergleich mit der Subkategorie *Korrektheit/Präzision* weniger Kommentare eingeteilt. Unter den vorhandenen Kommentaren finden sich Segmente, die auf einen globalen Eindruck des Spektrums von grammatischen Strukturen sowie deren Komplexität, d. h. einfache bzw. avancierte Strukturen (vgl. Beispiel 6.7) zurückzuführen sind. Die Bewertenden verwiesen jedoch auch auf das Spektrum spezifischer grammatischer Konstruktionen. Hierbei werden häufig Strukturen wie die

91 „Viele sprachliche Abweichungen gelten jedoch einfachen Verbformen, die sitzen sollten".
92 „Schafft die Nebensatzwortfolge zweimal, aber dennoch sind gewisse einfachere Verbfehler zu verzeichnen".
93 „Satzstellung unsicher (z. B. in Nebensätzen, Verbbeugung)".

Satzstellung im Nebensatz, die Verwendung verschiedener Verbformen und die allgemeine Variation im Satzbau erwähnt (vgl. Beispiele 6.8 und 6.9):

(6.7) Du använder avancerade konstruktioner och lyckas för det mesta riktigt bra med det.[94] (Cemu14-3-A, Lehrkraft)
(6.8) Strukturen: Teilweise angemessen (Satzstellung; der Passiv wird nicht beherrscht, ist aber Teil der Grammatik auf Niveau B1). (Pnmj1-5, GER-Bewert. 1)
(6.9) Eleven kan använda en varierad satsbyggnad.[95] (Kinv5-5-C, Lehrkraft)

Die Segmente der Subkategorie *Spektrum* enthalten oft Kommentare auf einer globalen Ebene, die übergreifend beschreiben, inwiefern ein einfaches oder avanciertes Spektrum von grammatischen Strukturen in den Schülertexten vorkommen (vgl. Beispiel 6.7). Häufig beziehen sich aber die Kommentare auch auf das Verwenden oder das Nicht-Verwenden spezifischer grammatische Konstruktionen. Dazu gehören auch Kommentare darüber, inwiefern eine gewisse grammatische Struktur dem sprachlichen Niveau entspricht (vgl. Beispiel 6.8). In diese Subkategorie gehören auch Kommentare über Variation im Satzbau (vgl. Beispiel 6.9). Zusammenfassend beziehen sich die Kommentare dieser Subkategorie folglich zum einen auf die Komplexität der vorhandenen formalen Strukturen. Zum anderen betreffen sie die Variation und Vielfalt formaler Strukturen in den Textproduktionen. Hierbei ergeben sich keine bedeutenden Unterschiede zwischen den Bewertergruppen.

Formale Strukturen: Orthographie

Ferner lassen sich in den Bewerterurteilen Hinweise auf die **Orthographie** in den Lernproduktionen finden. Aussagen über die orthographische Form werden von sowohl den schwedischen als auch den GER-Bewertenden gemacht. Einige Kommentare verweisen auf den globalen Eindruck der Orthographie (vgl. Beispiel 6.10), die meisten deuten aber auf spezifische orthographische Schwierigkeiten hin (vgl. Beispiele 6.11 und 6.12):

(6.10) viele Rechtschreiberfehler (Crpu19-4, GER-Bewert. 2)
(6.11) Substantiven med liten bokstav borde eleven komma ihåg med tanke på hur mycket jag tjatat om det.[96] (Hjbt5-3-E, Lehrkraft)
(6.12) Achtung: Substantive = groß, Kommas! (Sces17-4-A, Lehrkraft)

---

94 „Du verwendest avancierte Strukturen und es gelingt dir meistens sehr gut".
95 „Der Schüler/die Schülerin kann einen variierten Satzbau verwenden".
96 „Der Schüler/die Schülerin sollten sich an Substantive mit Kleinbuchstaben erinnern, angesichts dessen, wie oft ich darüber gemeckert habe".

Diese Kommentare beziehen sich hauptsächlich auf die Groß- und Kleinschreibung der Substantive (vgl. Beispiele 6.11 und 6.12) sowie die Beherrschung der Rechtschreibung (vgl. Beispiel 6.10). Eine geringere Anzahl von Kommentaren gilt der Verwendung von Satzzeichen (vgl. Beispiel 6.12).

Beide Bewertergruppen berücksichtigen in den Urteilen zu einem gewissen Grad die *Orthographie* in den Textproduktionen. Wenn sie die Beherrschung der Rechtschreibung kommentieren, wie an den Beispielen auch ersichtlich, wird diese häufig in negativen Worten erfasst (vgl. Beispiele 6.10–6.12), dies gilt sowohl für die globalen als auch für die lokalen Rechtschreibfehler. Kommentare zur Rechtschreibung scheinen somit hauptsächlich bei orthographischen Schwierigkeiten vorzukommen. In einigen Kommentaren der Lehrkräfte (vgl. Beispiel 6.11) können zudem auch Spuren des Unterrichts wahrgenommen werden, was darauf hindeuten könnte, dass die Lehrkraft das, was häufig im Unterricht behandelt worden ist, anders bewertet.

Aspekte, die in die Subkategorie *Orthographie* eingeordnet werden können, sind in diesem Zusammenhang interessant, da die Beherrschung der Orthographie zur linguistischen Kompetenz des GER gehört (vgl. Europarat 2001: 118 und hierzu Anhang 7), in den Kriterien der schwedischen Rahmenpläne für Sprachen jedoch nicht explizit erwähnt wird.[97] Immerhin wird Orthographie im Bewertungsraster des Goethe-Instituts (vgl. Anhang 12) als Beispiel eines Bewerteraspekts innerhalb der formalen Strukturen erwähnt.

Wortschatz: Differenziertheit

Die qualitative Analyse der Subkategorien zum ***Wortschatz*** zeigt sowohl Gemeinsamkeiten als auch Unterschiede mit der Analyse der formalen Strukturen auf. In den Bewerterurteilen zur Subkategorie ***Differenziertheit*** verweisen die Bewertenden häufig auf einen globalen Eindruck der lexikalischen Vielfalt. Diese Kommentare beziehen sich dabei hauptsächlich auf folgende Erscheinungen: die Differenziertheit des Wortschatzes (vgl. Beispiele 6.13 und 6.14) sowie die Adäquatheit des Wortschatzes (vgl. Beispiele 6.15 und 6.16):

(6.13) Otillräckligt ordförråd. Del 2: Lyckas få fram sin åsikt med ett mycket spartanskt och begränsat ordförråd.[98] (Rjrv2-5-F, Lehrkraft)

(6.14) Grundläggande ordförråd finns.[99] (Saig6-4-E, Lehrkraft)

---

97 Vgl. aber Hinweise zur Orthographie im Kommentarmaterial zu den Lehrplänen für *Moderna språk* (Skolverket 2011b: 15).

98 „Unzureichender Wortschatz. Teil 2: Schafft es, seine Meinung mit einem sehr spartanischen und begrenzten Wortschatz hervorzubringen“.

99 „Grundlegender Wortschatz vorhanden“.

(6.15) Ordförråd fungerar till uppgiften.[100] (Cemu14-3-C, ext. schwed. Bewert. 2)
(6.16) Wortschatz: Teilweise angemessen (mehr öffnen, Jugend […] „weder – noch“ nicht bekannt, gehört aber zum Wortschatz B1; leider). (Cemu14-3, GER-Bewert. 1)

Die große Mehrheit der Kommentare zur Differenziertheit des Wortschatzes sind entweder auf die allgemeine Vielfalt oder auf den Grad der Komplexität (vgl. Beispiel 6.14) zurückzuführen. Kommentiert wird zudem, wie Lernenden mit einem begrenzten Vokabular umgehen (vgl. Beispiel 6.13). Weniger Kommentare betreffen die Frage, inwieweit die jeweiligen Lernproduktionen ein adäquater Wortschatz beinhalten. Hierbei kommentieren häufiger die schwedischen Bewertenden und zwar dahingehend, inwiefern der Wortschatz einer spezifischen Aufgabe angemessen ist (vgl. Beispiel 6.15). Die GER-Bewertenden andererseits kommentieren gelegentlich, inwiefern der Wortschatz dem zu erwartenden GER-Niveau entspricht (vgl. Beispiel 6.16).

Wortschatz: idiomatische Ausdrücke

Die Subkategorie ***idiomatische Ausdrücke*** bezieht sich auf Phrasen, gebräuchliche Ausdrücke, Kollokationen oder feste Wendungen in den Lernproduktionen. Diese Kategorie wird hauptsächlich von den schwedischen Bewertenden verwendet, kaum aber von den GER-Bewertenden. Diese Kommentare enthalten Hinweise auf sowohl einen globalen Eindruck (vgl. Beispiel 6.17) als auch einzelne Kollokationen oder feste Wendungen in den Schülertexten (vgl. Beispiele 6.18 und 6.19):

(6.17) Det finns exempel på mer avancerade ord och uttryck: (*bin der Meinung*).[101] (Kasv3-5-A, Lehrkraft)
(6.18) „Mich nervt es“ med mera visar att man är språkligt litet mer än bara på en godkänd nivå.[102] (Crpu19-4-C, ext. schwed. Bewert. 1)
(6.19) Wortschatz: NB: „jdm um etwas bitten“ gehört zum B1 Wortschatz. (Kcku15-4, GER-Bewert. 1)

An den Beispielen wird deutlich, dass die Verwendung gebräuchlicher Ausdrücke aus dem Deutschen von den schwedischen Bewertenden sowohl positiv als auch negativ eingeschätzt wird. Die Kommentare der schwedischen Bewertenden bestehen fast im gleichem Maß aus Segmenten auf der globalen Ebene, die

---

100 „Der Wortschatz passend zur Aufgabe“.
101 „Es gibt Beispiele für fortgeschrittene Wörter und Ausdrücke“.
102 „‚Mich nervt es‘ u.v.m. zeigt, dass man sich sprachlich ein bisschen höher als auf einem ausreichenden Niveau befindet“.

generelle Aussagen über die Verwendung von Redewendungen und Phrasen im Hinblick auf Variation und Korrektheit treffen, wie aus solchen, die auf spezifische Ausdrücke in den Leistungen bezogen sind. Im einzigen Beispiel der GER-Bewertenden wird kommentiert, inwiefern die Lernenden Ausdrücke, die auf B1-Niveau liegen, bewältigen können (vgl. Beispiel 6.19). Auch in den Kommentaren der schwedischen Bewertenden sind aber Aussagen zum ausreichenden Niveau zu finden (vgl. Beispiel 6.18).

Wortschatz: Korrektheit/Präzision

Die Subkategorie ***Korrektheit/Präzision*** macht sowohl bei den schwedischen Bewertenden als auch bei den GER-Bewertenden den größten Anteil der Kommentare im Bereich des Wortschatzes aus. Unter diese Kategorie fallen Kommentare hinsichtlich der Präzision bei der Wortwahl, sowohl auf einer globalen Ebene (vgl. Beispiel 6.20) als auch lokal in den Textproduktionen (vgl. Beispiel 6.21). Eine Unsicherheit bezüglich der Präzision bei der Wortwahl bzw. der Beherrschung im Wortschatz sowie deren Einfluss auf die Verständlichkeit zeigt sich auch (vgl. Beispiele 6.22 und 6.23):

(6.20) Generellt en hel del fel ordvalsmässigt.[103] (Pnmj1-5-E, ext. schwed. Bewert. 2)
(6.21) Vokabel (*Erlebung*). (Slsk1-5-F, ext. schwed. Bewert. 1)
(6.22) Wortschatz: „praktizieren" in der Bedeutung „ein Praktikum machen" selten, aber verständlich; mehrere Fehlgriffe beeinträchtigen das Verständnis erheblich (nicht zu (so?) viel gemacht, Buchen gesahlt).(Cllu4-3, GER-Bewert. 1)
(6.23) Ibland ordval [som stör begripligheten] (treflig).[104] (Imns4-3-E, ext. schwed. Bewert. 2)

Wie an den Beispielen ersichtlich, sind die allermeisten Kommentare dieser Kategorie negativer Art. Bei einigen dieser Fehlgriffe bei der Wortschatzbeherrschung können Übertragungen aus dem Englischen oder Schwedischen wahrgenommen werden (vgl. Beispiele 6.22 und 6.23). Ebenso wie bei den formalen Strukturen setzen folglich die GER-Bewertenden wegen des Bewertungsrasters sehr häufig die Kommentare zur Korrektheit und Präzision des Wortschatzes mit der Verständlichkeit in Verbindung (vgl. Beispiel 6.22). Ein solcher Bezug tritt bei den schwedischen Bewertenden nicht so häufig auf. Es kommt zwar vor, dass die schwedischen Bewertenden den Einfluss der Beherrschung des Lexikons mit der Verständlichkeit verbinden (vgl. Beispiel 6.23), aber dies ist im analysierten Material eher eine Ausnahme.

---

103 „Generell viele Fehlgriffe bei der Wortwahl".
104 „Manchmal Wortfehler, die das Verständnis beeinträchtigen (treflig)".

Pauschale Beurteilung – Sprache

Die Segmente zur Kategorie ***pauschale Beurteilung – Sprache*** beziehen sich auf generalisierende Kommentare über die Sprache in den Schülertexten. Diese eher generellen Aussagen zur Sprache sind weder explizit auf den Wortschatz noch auf formale Strukturen zurückzuführen. Zu dieser Kategorie zählen Kommentare, die sich sowohl auf einen globalen Eindruck der Sprache als auch auf einzelne Textpassagen beziehen (vgl. Beispiele 6.24–6.27):

(6.24) Sammantaget svåra språkliga brister [som påverkar begripligheten].[105] (Geks8-3-F, ext. schwed. Bewert. 1)
(6.25) Vissa språkliga fel, [men detta stör inte & kommunikationen går fram]. Välformulerade texter.[106] (Sces175-4-A, ext. schwed. Bewert. 2)
(6.26) Stolpigt formulerat. Språket räcker knappt.[107] (Vwbg25-4-E, Lehrkraft)
(6.27) Die Meinungsäußerung ist wegen der sprachlichen Schwierigkeiten [schwer zu verstehen]. (Cswu1-3, GER-Bewert. 1)

Die Bewertenden kommentieren in ihren generellen Aussagen sowohl den globalen Eindruck einer begrenzten (Beispiele 6.24, 6.26 und 6.27) als auch einer gemischten oder fortgeschrittenen Sprachbeherrschung (vgl. Beispiel 6.25). Relativ häufig wird die pauschale Beurteilung der Sprache in Verbindung mit der Verständlichkeit in den Textproduktionen gesetzt (vgl. Beispiele 6.24, 6.25 und 6.27). Es ist zudem sehr deutlich, dass die schwedischen Bewertenden in höherem Ausmaß generelle Aussagen über die Sprache treffen, die auch größtenteils negativ sind (vgl. Beispiele 6.24 und 6.26). Generell ermitteln die schwedischen Bewertenden somit relativ häufig in ihren Urteilen eine pauschale Bewertung der Sprache, während Kommentare dieser Kategorie unter den GER-Bewertenden bis auf eine Ausnahme (vgl. Beispiel 6.27) nicht existieren.

6.3.2 Aspekte zur Verständlichkeit

Zur Hauptkategorie ***Verständlichkeit*** gehören die Subkategorien *Verständlichkeit – allgemein*, die einen Eindruck der Verständlichkeit aus der Perspektive der Bewertenden beinhaltet, und *Verwendung von Englisch oder Muttersprache*. Eine vertiefende Analyse der Bewerterkommentare hinsichtlich Aspekten zum Verständnis ergibt Gemeinsamkeiten, aber auch deutliche Diskrepanzen zwischen den schwedischen Bewertenden, siehe Tab. 20, und den GER-Bewertenden, wie Tab. 21 zeigt:

---

105 „Insgesamt viele sprachliche Defizite, die das Verständnis beeinträchtigen".
106 „Einige sprachliche Fehlgriffe, aber dies stört nicht & die Kommunikation schreitet voran. Gut formulierte Texte".
107 „Stolpernd formuliert. Die Sprache reicht kaum aus".

Tab. 20: *Verteilung der Bewerterkommentare der schwedischen Bewertenden (N = 180) auf die Verständlichkeit (Anzahl der Segmente)*

Verständlichkeit	*negativ*	*gemischt*	*positiv*	***Gesamt***
– allgemein	53	16	30	**99**
– Verwendung von Eng/L1	22	0	1	**23**
Verständlichkeit – gesamt	**75**	**16**	**31**	**122**

Tab. 21: *Verteilung der Bewerterkommentare der GER-Bewertenden (N = 120) auf die Verständlichkeit (Anzahl der Segmente)*

Verständlichkeit	*negativ*	*gemischt*	*positiv*	***Gesamt***
– allgemein	10	51	48	**109**
– Verwendung von Eng/L1	3	0	0	**3**
Verständlichkeit – gesamt	**13**	**51**	**48**	**112**

Kommentare, die in die Kategorie *Verständlichkeit* eingeordnet werden können, finden sich in beiden Bewertergruppen. Bezüglich der Subkategorien ergeben sich dennoch Unterschiede zwischen den Bewertergruppen. Zum einen kommentieren die schwedischen Bewertenden die *Verwendung von Schwedisch oder Englisch*, während die GER-Bewertenden dies in ihren Bewerterurteilen kaum berücksichtigen. Zum anderen überwiegen, auch wenn beide Bewertergruppen die *Verständlichkeit – allgemein* beachten, unter den schwedischen Bewertenden die negativ eingeschätzten Kommentare. Die schwedischen Bewertenden kommentieren dementsprechend, wie z. B. auch bei orthographischen Schwierigkeiten, häufiger das Verständnis, wenn die Texte in diesem Bereich Probleme aufweisen. Die GER-Bewertenden geben hingegen kaum negative Kommentare in dieser Hinsicht. In ihren Bewerterurteilen überwiegen in dieser Kategorie stattdessen die gemischten und die positiven Kommentare.

Verständlichkeit: allgemein

Zur Subkategorie ***Verständlichkeit – allgemein*** gehören Hinweise auf die globale Klarheit der Darstellung (vgl. Beispiel 6.28). Die Bewertenden setzen auch Aspekte zum Verständnis mit anderen Phänomenen in den Textproduktionen in Verbindung (vgl. Beispiele 6.29–6.31):

(6.28) Men texterna är ändå begripliga.[108] (Hjbt5-3-E, ext. schwed. Bewert. 2)
(6.29) Svårförståeligt [pga inkorrekt meningsbyggnad.][109] (Saig6-4-F, ext. schwed. Bewert. 1)
(6.30) Strukturen: Mehrere Fehlgriffe beeinträchtigen das Verständnis teilweise (Seit Ich vor die letzte Woche krank bin; ich hopfe da du mir Entschuldigung kann). (Vmeg5-3, GER-Bewert. 1)
(6.31) Några obegripliga ord.[110] (Vedg-3-E, Lehrkraft)

Auch wenn Kommentare, die ausschließlich einen globalen Eindruck der Verständlichkeit geben, im Material vorkommen (vgl. Beispiel 6.28), sind diese nur in den Kommentaren der schwedischen Bewertenden zu finden. Häufiger wird von sowohl den schwedischen als auch den GER-Bewertenden die Beeinflussung des Verständnisses durch sprachliche Schwierigkeiten kommentiert. Diese beziehen sich entweder auf den ganzen Text (vgl. Beispiel 6.29) oder auf Fehlgriffe einzelner Textpassagen, Phrasen oder Wörter (vgl. Beispiele 6.30 und 6.31).

Kommentare zur Verständlichkeit beziehen sich häufig, wie auch aus den Beispielen oben ersichtlich, auf grammatische oder lexikalische Fehlgriffe in den Schülerleistungen (z. B. Beispiele 6.29 und 6.31). Im Datensatz lässt sich aber ein deutlicher Unterschied zwischen den beiden Bewertergruppen beobachten: GER-Bewertende setzen häufiger als die schwedischen Bewertenden Fehlgriffe in den Bereichen formaler Strukturen oder Wortschatz mit der Verständlichkeit des Schülertextes in Verbindung (vgl. Beispiel 6.30). Dies wird auch deutlich, wenn zwei Bewertende im Hinblick auf die Verständlichkeit die gleichen Textpassagen in den Textproduktionen kommentieren, siehe Beispiele 6.32 und 6.33:

(6.32) om än på vissa ställen oklart, ex. „beide Bücher magst und ein Morgenperson bist"[111] (Vnjg2-3-C, ext. schwed. Bewert. 1)
(6.33) Wortschatz: Fehlgriffe beeinträchtigen das Verständnis nicht (… dass man ins Zeit kommen müss; ich weiß, dass du, Jürgen, beide Bücher magst und ein Morgenperson bist). (Vnjg2-3, GER-Bewert. 1)

Während der schwedische Bewertende eine generellere Aussage trifft, setzt der GER-Bewertende hier Fehlgriffe im Bereich des Wortschatzes mit der

108 „Aber die Texte sind dennoch verständlich".
109 „Schwer verständlich wegen inkorrektem Satzbau".
110 „Einige unverständliche Wörter".
111 „Wenn auch gelegentlich unklar, z. B. ‚beide Bücher magst und ein Morgenperson bist' ".

Verständlichkeit in Verbindung und findet, dass diese das Verständnis nicht beeinträchtigt.

In den Kommentaren können auch individuelle Unterschiede zwischen den Bewertenden wahrgenommen werden. Auffällig hierbei sind die Kommentare des ersten schwedischen externen Bewertenden, der häufig auf sog. „zerstörende Fehler“ hinweist (vgl. Beispiel 6.34):

(6.34) Inledningen bra och förståelig. Så följer några svårtydda meningar. De är s.k. „förstörande fel“, dvs en tyskspråkig person skulle ha svårt att förstå.[112] (Vedg3-3-C, ext. schwed. Bewert. 1)

Der Ausdruck „zerstörende Fehler“ wird in diesem Beispiel als Unklarheiten erklärt, die einer deutschsprachigen Person ohne Schwedischkenntnisse (oder Kenntnisse des Englischen) Schwierigkeiten beim Verstehen bereiten dürften. Gerade der Ausdruck „zerstörende Fehler“ stammt wahrscheinlich aus den generellen Bewertungsanweisungen zum nationalen Testmaterial für die Fremdsprachen in Schweden. In diesen Anweisungen wird deutlich, dass in den Leistungen Stärke vor Schwäche erhoben werden sollten und dabei sollte zwischen Fehlgriffen, die das Verständnis beeinträchtigen, und Fehlgriffen, die die Kommunikation stören könnten, unterschieden werden (vgl. Erickson 2020b). Das Geschriebene sollte nach den Anweisungen für eine deutschsprachige Person verständlich sein und einzelne Fehlgriffe im Text dürfen die Kommunikation stören, aber eben nicht zerstören. Dies bedeutet oft auch, dass Elemente aus dem Schwedischen oder aus anderen Sprachen, der Fokus für den folgenden Teilabschnitt, vermieden werden sollten.[113]

Verständlichkeit: Verwendung von Englisch/L1

Die Subkategorie zur ***Verwendung von Englisch und der Muttersprache*** bezieht sich auf das Auftreten von Textpassagen, die Bezug auf Englisch, Schwedisch oder beide Sprachen nehmen. Es handelt sich dabei manchmal um einen globalen Eindruck des Einflusses von englischen oder schwedischen Elementen (vgl. Beispiel 6.35). Häufiger wird der Einfluss anderer Sprachen spezifiziert, wie u. a. die Übertragung von Texteinheiten oder Phrasen (vgl. Beispiele 6.36

112 „Einleitung gut und verständlich. Dann folgen einige schwer verständliche Sätze. Diese sind sog. ‚zerstörende Fehler‘, d. h. eine deutschsprachige Person würde Schwierigkeiten zu verstehen haben“.

113 Vgl. hierzu z. B. die generellen Anweisungen des nationalen Prüfungsmaterials für die Fremdsprachen Deutsch, Französisch und Spanisch (Skolverket 2021b).

und 6.37), syntaktischen oder morphologischen Phänomenen (vgl. Beispiel 6.38) sowie einzelnen Vokabeln (vgl. Beispiel 6.39):

(6.35) Blandar in engelska.[114] (Shfg3-4-F, ext, schwed. Bewert. 1)
(6.36) Englisch: vielen Geld gemacht. (Ghhs4-3-D, ext, schwed. Bewert. 1)
(6.37) Wortschatz: mehrere Fehlgriffe beeinflussen das Verständnis (treflig, alles alena; am 24/7–24/7 wird im Deutschen selten verwendet, eher „rund um die Uhr“, „24 Stunden). (Imns4-3, GER-Bewert. 1)
(6.38) Satzlösungen erinnern an Schwedisch und sind mitunter umständlich. (Sess13-4-C, Lehrkraft)
(6.39) med en del svenska ord[115] (Imns4-3-F, Lehrkraft)

Bezugnahmen auf Englisch oder Schwedisch in den Bewerterurteilen sind fast im gleichen Ausmaß auf die Sprachen verteilt, Belegstellen auf die Verwendung von Englisch überwiegen aber. Hierbei können zudem interessante Gemeinsamkeiten und Unterschiede zwischen den Sprachen wahrgenommen werden. Die meisten Belegstellen in den Kommentaren, die auf die Verwendung von Schwedisch oder Englisch zurückzuführen sind, beziehen sich auf das Lexikon (z. B. Beispiel 6.39). Die Verwendung von Englisch bezieht sich aber in den Urteilen häufig auch auf Phrasen (z. B. Beispiel 6.36).

Unter den Kommentaren zum Gebrauch von Schwedisch ergibt sich hingegen eine Tendenz, Belegstellen bezüglich der Verwendung des schwedischen Satzbaus zu finden (vgl. Beispiel 6.38) – ein Aspekt, der überhaupt nicht in den Kommentaren über Englisch vorzufinden ist. Übertragungen im Hinblick auf syntaktische Phänomene scheinen dementsprechend in den Textproduktionen eher aus dem Schwedischen zu kommen. Hinweise darauf, dass die Bewertenden in den Textproduktionen andere Sprachen als Schwedisch oder Englisch gefunden haben, sind in den Bewerterurteilen nicht vorhanden.

Die schwedischen Bewertenden kommentieren generell im Vergleich zu den GER-Bewertenden in höherem Grad die Verwendung von Wörtern aus dem Englischen oder Schwedischen (vgl. Tab. 20 und Tab. 21). Beide Bewertergruppen setzten jedoch häufig die Verwendung von Englisch und Schwedisch in Relation zur Verständlichkeit des Textes (vgl. Beispiele 6.40 und 6.41):

(6.40) Wortschatz: („scary“ ist der einzige Fehlgriffe, der das Verständnis für Leser, die kein Englisch verstehen, beeinträchtig). (Kiiu2-4, GER-Bewert. 1)
(6.41) Svenska. […] Svårt att förstå, gör många ordfel.[116] (Kefu5-4-E, ext. schwed. Bewert. 1)

114 „Mischt Englisch bei“.
115 „Mit einigen schwedischen Wörtern“.
116 „Schwedisch. […] Schwer zu verstehen, macht viele Wortfehler“. Der Kommentar „Schwedisch“ bezieht sich hier auf das schwedische Wort „minne“

Die Verwendung von Englisch und/oder Schwedisch wird in den Bewerterurteilen bis auf eine Ausnahme ausschließlich negativ bewertet. Bei dem einzigen positiv bewertenden Beispiel handelt es sich um eine Lehrkraft, die das Vermeiden englischer und schwedischer Wörter als eine gute Strategie versteht.

6.3.3 Aspekte zur Aufgabenerfüllung

Die Kommentare in der Kategorie ***Aufgabenerfüllung***, sind weiter in Subkategorien über die inhaltliche Erfüllung (*Aufgabenerfüllung – Inhalt*) bzw. die umfängliche Erfüllung (*Aufgabenerfüllung – Textlänge*) unterteilt. In den Tabellen unten wird die Häufigkeit der positiv, gemischt und negativ kodierten Segmente in den jeweiligen Subkategorien dargestellt. Die Ergebnisse werden getrennt angegeben, vgl. Tab. 22 für die schwedischen Bewertenden und Tab. 23 für die GER-Bewertenden:

Tab. 22: *Verteilung der Bewerterkommentare der schwedischen Bewertenden (N = 180) auf die Aufgabenerfüllung (Anzahl der Segmente)*

Aufgabenerfüllung	*negativ*	*gemischt*	*positiv*	***Gesamt***
– Inhalt	28	17	52	**97**
– Textlänge	45	1	4	**50**
Aufgabenerfüllung – gesamt	**73**	**18**	**56**	**147**

Tab. 23: *Verteilung der Bewerterkommentare der GER-Bewertenden (N = 120) auf die Aufgabenerfüllung (Anzahl der Segmente)*

Aufgabenerfüllung	*negativ*	*gemischt*	*positiv*	***Gesamt***
– Inhalt	6	55	57	**118**
– Textlänge	30	5	48	**83**
Aufgabenerfüllung – gesamt	**36**	**60**	**105**	**201**

Die Analyse zur *Aufgabenerfüllung* ergibt, dass Kommentare zu inhaltlichen Aspekten im Vergleich zu Aussagen über die Textmenge von beiden Bewertergruppen häufiger durchgeführt werden. Ferner können auch in der Analyse zur inhaltlichen Aufgabenerfüllung Unterschiede zwischen den Bewertergruppen wahrgenommen werden. Hierbei sind deutliche Diskrepanzen bezüglich der

[deutsch: Erinnerung] im Text. In der Schülerleistung steht: „Ein Foto ist ein ‚minne' für dich".

Verteilung von positiven, gemischten bzw. negativen Kommentaren zu finden, vor allem hinsichtlich der inhaltlichen Aufgabenerfüllung. Auffallend ist dabei, dass die Kommentare der GER-Bewertenden in den allermeisten Fällen gemischt oder positiv sind. Auch die schwedischen Bewertenden haben großteils positive Kommentare, formulieren aber häufiger als die GER-Bewertenden inhaltliche Aspekte in negativen Worten.

Des Weiteren ergeben sich Unterschiede im Hinblick auf die Textlänge. Kommentare zur Textlänge werden häufiger von den GER-Bewertenden als von schwedischen Bewertenden gegeben. In den Bewerterurteilen der GER-Bewertenden überwiegen die positiven Kommentare, wobei dennoch etwa ein Drittel der Kommentare eher negativ sind und der nicht erreichten Mindestanzahl von Wörtern gelten. Hierbei scheint die explizite Erwähnung der erfüllten bzw. nicht-erfüllten Textlänge im Bewertungsraster der GER-Bewertenden eine Rolle zu spielen: Die Textlänge wird, offensichtlich wegen der Einwirkung des Bewertungsrasters, von den GER-Bewertenden erwähnt, auch wenn die Anzahl der Wörter umfänglich angemessen ist. Die große Mehrheit der Kommentare der schwedischen Bewertenden bezüglich der Textlänge ist negativ. Sie kommentieren demzufolge diesen Aspekt eher, wenn er nicht erfüllt ist. Unter den schwedischen Bewertenden sind zudem weitere Unterschiede zu erkennen, siehe Tab. 24:

Tab. 24: *Verteilung der Bewerterkommentare der schwedischen Lehrkräfte bzw. der schwedischen externen Bewertenden (N = 180) auf die Aufgabenerfüllung (Anzahl der Segmente)*

Aufgabenerfüllung	*Gruppe der Lehrkräfte*	*ext. schwed. Bewert. 1*	*ext. schwed. Bewert. 2*	***Gesamt***
– Inhalt	25	23	49	**97**
– Textlänge	8	30	12	**50**
Aufgabenerfüllung – gesamt	**33**	**53**	**61**	**147**

Die schwedischen Deutschlehrkräfte scheinen im Vergleich zu den zwei externen schwedischen Bewertenden in geringerem Ausmaß Aspekte der Aufgabenerfüllung in den Textproduktionen zu beachten. Auch zwischen den beiden externen schwedischen Bewertenden ergeben sich individuelle Tendenzen: Während die/der erste externe Bewertende eine Tendenz hat, zu einem höheren Grad die Textlänge zu berücksichtigen, beachtet die/der zweite externe Bewertende in höherem Ausmaß die inhaltlichen Aspekte.

Aufgabenerfüllung: Inhalt

Die Kommentare zur ***inhaltlichen Aufgabenerfüllung*** beziehen sich häufig auf einen globalen Eindruck der inhaltlichen Erfüllung, entweder in den gesamten Textproduktionen (vgl. Beispiel 6.42) oder in den Teilaufgaben (vgl. Beispiel 6.43):

(6.42) Inhalt begrenzt. (Sons4-3-E, Lehrkraft)
(6.43) Fattas en del av uppgiften för innehållet.[117] (Sons4-3-F, ext. schwed. Bewert. 1)

Zur inhaltlichen Aufgabenerfüllung zählen darüber hinaus Kommentare über die inhaltliche Qualität einzelner Sprachfunktionen (vgl. Beispiel 6.44) oder über die Erfüllung bzw. Nicht-Erfüllung der in der Aufgabe nachgefragten Sprachfunktionen (vgl. Beispiel 6.45). Belegstellen zur inhaltlichen Erfüllung umfassen zudem Kommentare darüber, inwiefern eine der Teilaufgaben überhaupt behandelt worden ist (vgl. Beispiel 6.46):

(6.44) Förklarar redigt varför uppgiften ej har hunnit göras.[118] (Kbtu25-4-B, ext. schwed. Bewert. 1)
(6.45) Alla tre uppgifter genomförs enligt instruktion (undantag att en träffpunkt ej föreslås).[119] (Imls9-4-E, ext. schwed. Bewert. 2)
(6.46) Erfüllung: Teilaufgabe nicht gelöst. (Örkl1-3, GER-Bewert. 2)

Wie an den Beispielen ersichtlich wird, beinhalten die Kommentare sowohl qualitative als auch quantitative Aspekte der inhaltlichen Anforderungen. Die Angaben zur Subkategorie *Inhalt* decken somit sowohl eine qualitative Perspektive, wo sich der Bewertende darauf bezieht, *wie gut* die Schülerinnen und Schüler die Aufgabe inhaltlich erfüllen, als auch eine quantitative Perspektive ab, wo berücksichtigt wird, inwiefern die jeweiligen inhaltlichen Anforderungen der Prüfung *überhaupt erfüllt* sind. Generell scheinen sowohl die schwedischen Bewertenden als auch die GER-Bewertenden die allgemeine Erfüllung des Inhaltes sowie die inhaltliche Bewältigung der angeforderten Sprachfunktionen zu beachten (vgl. Beispiel 6.44). Nachgefragte Sprachfunktionen in der Aufgabe sind z. B. eine Entschuldigung oder einen Vorschlag für ein Treffen zu formulieren (vgl. Beispiel 6.45).

Auch Kommentare, die auf eine nicht gelöste Teilaufgabe hinweisen, sind bei beiden Bewertergruppen zu finden (vgl. Beispiel 6.46). Letztere finden sich,

117 „Ein Teil der Aufgabe fehlt für die inhaltliche Erfüllung".
118 „Erklärt richtig, warum die Aufgabe nicht gemacht worden ist".
119 „Alle drei Aufgaben werden nach den Anweisungen ausgeführt (Ausnahme, dass ein Vorschlag für ein Treffen fehlt)".

was kaum überrascht, fast ausschließlich in Beurteilungen hinsichtlich nicht-ausreichender Textproduktionen. Auffallend sind aber die wenigen Kommentare der Gruppe der schwedischen Deutschlehrkräfte im Hinblick auf die inhaltliche Erfüllung. Die Kommentare der Lehrkräfte beziehen sich häufiger auf einen allgemeinen Eindruck des Inhalts oder darauf, inwiefern die Aufgabe gelöst ist. Deutlich seltener betreffen ihre Kommentare die Erfüllung der in der Aufgabe angeforderten Sprachfunktionen, wie die Aufforderung, eine Entschuldigung zu schreiben.

Aufgabenerfüllung: Textlänge

Die Subkategorie zur umfänglichen Aufgabenerfüllung (die ***Textlänge***) zählt zu den quantitativen Komponenten, die sich in Bewertungskriterien nicht immer wiederfinden. Die Länge der Texte kann jedoch, zumindest in diesem Fall, als ein Teil der Aufgabenerfüllung aufgefasst werden, da eine Mindestwortzahl für die jeweiligen Aufgaben angegeben wurde (vgl. Anhang 9). Die Kommentare der Bewertenden zur Textlänge beziehen sich darauf, wie die Lernenden sich an diese Mindestwortzahl gehalten haben. Am häufigsten kommentiert werden die Fälle, in denen die Lernenden die geforderte Wortanzahl nicht erreicht haben (vgl. Beispiele 6.47 und 6.48). Auch Kommentare hinsichtlich umfänglich angemessener Textproduktionen sind aber in den Urteilen zu finden (vgl. Beispiele 6.49 und 6.50):

(6.47) Erfüllung: weniger als 50% der geforderten Wortanzahl. […] Erfüllung: sehr kurz. (Hmlt2-3, GER-Bewert. 2)

(6.48) Något för kortfattad [och ger därför ej tillräckligt med kommunikativt innehåll].[120] (Vjrg24-4-D, ext. Schwed. Bewert. 1)

(6.49) Anstränger sig dock att hålla begärd längd på delarna.[121] (Ilms9-4-F, ext. schwed. Bewert. 1)

(6.50) Erfüllung: Alle 3 Sprachfunktionen umfänglich (gerade noch) angemessen behandelt. (Vedg3-3, GER 1)

Häufig wird im Hinblick auf die Textlänge in den Kommentaren die Erfüllung der Mindestwortzahl beachtet. Einige der Kommentare beziehen sich, wie an den Beispielen ersichtlich, nicht nur auf die genaue Wortanzahl, sondern auch darauf, inwiefern die Textproduktionen zugleich auch den kommunikativen Inhalt der Aufgabe erfüllt (vgl. Beispiel 6.48). Kommentare zur Textlänge finden sich im höheren Grad in den Urteilen der GER-Bewertenden.

120 „Etwas kurz und bietet deswegen nicht genügend kommunikativen Inhalt“.

121 „Strengt sich jedoch an, die angeforderte Länge der Teile zu halten“.

6.3.4 Aspekte zur Angemessenheit

Die Kategorie ***Angemessenheit*** ist in die Subkategorien *Kohärenz*, *soziokulturelle Angemessenheit*, *Textaufbau* und *Textsorte* gegliedert. In den Tabellen unten wird die Verteilung der Kommentare der schwedischen Bewertenden (vgl. Tab. 25) bzw. der GER-Bewertenden (vgl. Tab. 26) dargestellt:

Tab. 25: *Verteilung der Bewerterkommentare der schwedischen Bewertenden (N = 180) auf Angemessenheit (Anzahl der Segmente)*

Angemessenheit	*negativ*	*gemischt*	*positiv*	***Gesamt***
– Kohärenz	14	2	6	**22**
– soziokulturell	29	11	35	**75**
– Textaufbau	8	0	16	**24**
– Textsorte	12	2	21	**35**
Angemessenheit – gesamt	**63**	**15**	**78**	**156**

Tab. 26: *Verteilung der Bewerterkommentare der GER-Bewertenden (N = 120) auf Angemessenheit (Anzahl der Segmente)*

Angemessenheit	*negativ*	*gemischt*	*positiv*	***Gesamt***
– Kohärenz	9	51	45	105
– soziokulturell	9	16	38	63
– Textaufbau	13	38	47	98
– Textsorte	25	1	1	27
Angemessenheit – gesamt	**56**	**106**	**131**	**293**

Hierbei ergibt sich eine Diskrepanz in der Verteilung. Insgesamt geben GER-Bewertenden wesentlich häufiger Kommentare zur Angemessenheit, 293 gegen 156 Segmente. Dies ist der Fall, obwohl die Gesamtanzahl der Bewertungen der GER-Bewertenden (120) niedriger ist als die Gesamtanzahl der schwedischen Bewertungen (180). Wie zudem aus den Tabellen hervorgeht, sind Aspekte, die in die jeweiligen Subkategorien zur Angemessenheit eingeordnet werden können, von den Bewertergruppen unterschiedlich berücksichtigt worden. Während die schwedischen Bewertenden in etwas höherem Ausmaß als die GER-Bewertenden die soziokulturelle Angemessenheit kommentieren, beachten die GER-Bewertenden hingegen häufiger Merkmale der Textorganisation, d. h. Aspekte zur Kohärenz und zum Textaufbau.

Ferner können auch Diskrepanzen im Hinblick auf die Verteilung in positive, gemischte bzw. negative Kommentare der Bewertergruppen bezüglich der

Aufgabenerfüllung verortet werden. Unter den schwedischen Bewertenden überwiegen insgesamt knapp die positiv wertenden Kommentare, wobei die negativ wertenden Kommentare einen fast gleich großen Anteil ausmachen. Der von den schwedischen Bewertenden am wenigsten beachtete Aspekt, *die Kohärenz*, wird häufiger negativ bewertet. Auch wenn die Kommentare der GER-Bewertenden im Einklang mit den schwedischen Bewertenden überwiegend positiv sind, bestehen sie, dem Muster anderer Kategorien folgend, in größerem Ausmaß als bei den schwedischen Bewertenden aus Kommentaren gemischter oder positiver Art. Dies gilt insbesondere für die Subkategorien *Kohärenz*, *soziokulturelle Angemessenheit* bzw. *Textaufbau*.

Die Kommentare der GER-Bewertenden zur *Textsorte* sind hingegen fast ausschließlich, und im Unterschied zu den schwedischen Bewertenden, negativ formuliert. Da die Kommentare zur Textsorte durch die GER-Bewertenden so deutlich negativ eingeschätzt werden, scheint die Anpassung an die Textsorte von den GER-Bewertenden nur bei Schwierigkeiten berücksichtigt zu werden. Die schwedischen Bewertenden vergeben hingegen Kommentare, die sowohl positiv als auch negativ wertend eingeordnet werden können.

Angemessenheit: Kohärenz

Die Kommentare zur Subkategorie ***Kohärenz*** beziehen sich überwiegend auf einen globalen Eindruck der Kohärenz (vgl. Beispiel 6.51), wobei gelegentlich deren Einfluss auf die Verständlichkeit erwähnt wird (vgl. Beispiel 6.52). Zudem finden sich in dieser Kategorie Kommentare unlogischer Satzverbindungen (vgl. Beispiel 6.53) und zur Verwendung von Konnektoren in den Textproduktionen (vgl. Beispiel 6.54):

(6.51) Framställningen är relativt sammanhängande.[122] (Gols6-4-A, ext. schwed. Bewert. 2)

(6.52) Din text är för osammanhängande [för att bli tillräckligt begriplig för att nå kunskapskraven för betyget E].[123] (Örkl-3-F, Lehrkraft)

(6.53) Ej sammanhängande. I flera fall hänger texten inte ihop logiskt (Meine opa liesen Bücher…).[124] (Crmu17-4-F, ext. schwed. Bewert. 2)

(6.54) Kohärenz: Überwiegend angemessen (Gute Verwendung von Konjunktionen „obwohl“, „so dass“). (Vmeg5-3, GER-Bewert. 1)

---

122 „Die Textproduktion ist relativ kohärent“.

123 „Dein Text ist allzu inkohärent um verständlich genug für die Anforderungen der Note E zu werden“.

124 „Inkohärent. In vielen Fällen hängt der Text nicht logisch zusammen (Meine opa liesen Bücher …)“.

In allen Beispielen wird deutlich, dass sich die Kommentare zur Kohärenz für sowohl die schwedischen Bewertenden als auch die GER-Bewertenden auf die globale Ebene in den Textproduktionen beziehen. In der geringen Anzahl der Kommentare von den schwedischen Bewertenden überwiegen Kommentare darüber, inwiefern die Texte zusammenhängend sind oder nicht (vgl. Beispiele 6.51–6.53). Nur in wenigen Fällen kommentieren die schwedischen Bewertenden die Kohärenz auch im Hinblick auf einzelne Bindewörter oder auf unlogische Satzverbindungen. Die Kommentare der GER-Bewertenden sind deutlich vom Bewertungsraster beeinflusst, da sie häufig einen globalen Eindruck zur Verknüpfung von Sätzen und Satzteilen, manchmal mit zusätzlichen Ergänzungen, wiedergeben (vgl. Beispiel 6.54).

Angemessenheit: soziokulturell

Unter den Kommentaren zur ***soziokulturellen Angemessenheit*** können in den Bewerterurteilen gewisse Unterschiede, aber dennoch auch Gemeinsamkeiten zwischen den Bewertergruppen bemerkt werden. Die Kommentare hierzu beziehen sich oft auf einen globalen Eindruck der soziokulturellen Angemessenheit, d. h. der allgemeinen Fähigkeit, sich sprachlich in formellen bzw. informellen Situationen anzupassen (vgl. Beispiel 6.55). In dieser Kategorie finden sich zudem spezifische Kommentare u. a. zur Verwendung von Anredeformen (vgl. Beispiel 6.56) sowie zu einem der Situation angepassten Register (vgl. Bespiel 6.57). Zu letzteren gehört auch der Gebrauch geeigneter Grußformeln, die in der gegebenen Situation partneradäquat sind (vgl. Beispiel 6.58):

(6.55) Mitteilung soziokulturell angemessen. (Eles2-5, GER-Bewert. 2)
(6.56) Hittar inte rätt tilltal – du och ni-tilltal sammanblandat.[125] (Rjrv2-5-F, Lehrkraft)
(6.57) Hälsning „Hallo“, Stil: „Wie geht’s, mann?“.[126] (Vedg3-3-F, Lehrkraft)
(6.58) Inga riktiga formella inlednings- och avslutningsfraser. Skriver dock på ett bra formellt sätt i brevet.[127] (Hjjg-5-C, ext. schwed. Bewert. 1)

Wie an den Beispielen ersichtlich wird, bestehen die Kommentare der schwedischen Bewertenden oft aus konkreten Textbeispielen, z. B. der Verwendung von Anredepronomen oder Grußformeln. Die GER-Bewertenden geben häufiger als die schwedischen Bewertenden einen Gesamteindruck der soziokulturellen

125 „Findet nicht die richtige Anrede – Mischen der Du- und Sie-Anrede“.
126 „Gruß ‚hallo‘, Stil: ‚Wie geht’s, mann?‘ “.
127 „Keine richtigen formellen Einleitungs- und Abschlussphrasen. Schreibt aber im Brief formell richtig“.

Angemessenheit, wobei sie aber auch häufig Ergänzungen bezüglich der Anredeformel vornehmen.

Angemessenheit: Textaufbau

Kommentare, die in die Subkategorie ***Textaufbau*** einzuordnen sind, kommen in höherem Grad bei den GER-Bewertenden vor. Sie beziehen sich in den Bewerterurteilen auf eine klare Struktur in den Textproduktionen (vgl. Beispiele 6.59 und 6.60). Hierzu gehören zudem Kommentare hinsichtlich Einleitung und Abschluss (vgl. Beispiele 6.61 und 6.62) sowie zu einer Unterteilung in Abschnitte (vgl. Beispiel 6.63):

(6.59) till disposition bra gjort.[128] (Kasv3-5-A, ext. schwed. Bewert. 1)
(6.60) Textaufbau überwiegend angemessen. (Cemu14-3, GER-Bewert. 2)
(6.61) (Einleitung, und Schluss […] fehlen). (Cllu4-3, GER-Bewert. 1)
(6.62) Ett abrupt slut på del tre [men i övrigt texter som ändå fungerar].[129] (Sing1-4-E, ext. schwed. Bewert. 2)
(6.63) Styckeindelning saknas.[130] (Pnmj1-5-F, Lehrkraft)

Auch an diesen Beispielen zeigen sich die zum überwiegenden Teil positiven Belege im Hinblick auf den Textaufbau und dies gilt sowohl für die schwedischen Bewertenden als auch für die GER-Bewertenden. Unter den Kommentaren der GER-Bewertenden dominieren verstärkt Äußerungen zur globalen Ebene, wobei diese zum Teil den vorgelegten Formulierungen aus dem Bewertungsraster folgen (vgl. Beispiel 6.60). Wenn die GER-Bewertenden aber zusätzliche Ergänzungen hinzufügen, beinhalten diese häufig Kommentare über die Einleitung bzw. den Abschluss in den Textproduktionen (vgl. Beispiel 6.61). Die schwedischen Bewertenden kommentieren hinsichtlich des Textaufbaus vor allem, inwiefern Einleitung oder Abschluss in den Leistungen vorhanden sind oder im Zusammenhang gut funktionieren (vgl. Beispiel 6.62). Zusammenfassend lässt sich erkennen, dass Kommentare beider Bewertergruppen zum Textaufbau auf die Qualität oder das Vorhandensein der Einleitung bzw. des Abschlusses verweisen. Aussagen zur Einleitung in Abschnitte kommen in den Kommentaren der beiden Bewertergruppen selten vor.

128 „Disposition ist gut".
129 „Ein abruptes Ende von Teil drei aber ansonsten funktionierende Texte".
130 „Unterteilung in Abschnitte fehlt".

Angemessenheit: Textsorte

Die Bewertenden kommentieren in den Urteilen insgesamt in etwas geringerem Ausmaß Aspekte der ***Textsorte***. Diese Kommentare beziehen sich oft auf einen globalen Eindruck der jeweiligen Textsorten (vgl. Beispiel 6.64), aber in diese Subkategorie gehören auch Aussagen zur Erfüllung bzw. Nicht-Erfüllung der Konventionen einer spezifischen Textsorte (vgl. Beispiele 6.65–6.67):

(6.64) Textsorte durchgängig umgesetzt. (Cemu14-3, GER-Bewert. 2)

(6.65) Dessvärre har du inte följt instruktionerna. Det skulle ju vara ett mail och då börjar man med „Hallo + namn“ eller „Liebe(r) + namn“.[131] (Cllu4-3-F, Lehrkraft)

(6.66) Bestandteile eine Mail wie Anrede Einleitung, Schluss, fehlen. […] kein Gruss am Schluss. (Crmu17-4, GER-Bewert. 2)

(6.67) Eleven har följt konventionerna för de olika texttyperna och anpassat språket och innehållet efter texttyp på ett bra och övertygande sätt. I uppgift 1 och uppgift 3 kan inlednings- och hälsningsfrasen förbättras. [*am Textrand notiert: „Viele Grüße“ anstatt des Wortes „Danke“. Eigene Ergänzung, M.H.R*].[132] (Kcku15-4-A, Lehrkraft)

Die Verwendung textsortenspezifischer Merkmale kann als die Fähigkeit zur Anpassungen an das Textgenre in einem deutschen Kontext verstanden werden und liegt somit der soziokulturellen Angemessenheit nahe. Die Kommentare zur Textsorte finden sich im Material ausschließlich bei der ersten und dritten Aufgabe des Tests, die das Schreiben einer formellen bzw. informellen E-Mail beinhalten. Wie auch an den Beispielen deutlich, gehören in diese Subkategorie häufig Kommentare über Gestaltungskonventionen der Sprachgemeinschaft zur Textsorte, in diesem Fall die Konventionen einer informellen bzw. formellen E-Mail. Es geht hierbei oft darum, inwiefern die Anreden einen Namen (z. B. Beispiel 6.65) oder andere wichtige Bestandteile wie zur Textsorte passende Grußformeln enthalten (vgl. Beispiel 6.67).

---

131 „Leider hast du die Anweisungen nicht befolgt. Es sollte eine E-Mail sein und dann fängt man mit ‚Hallo + Name‘ oder ‚Liebe(r) + Name‘ an“.

132 „Der Schüler/die Schülerin hat die Konventionen für die verschiedenen Texttypen befolgt und Sprache und Inhalt gut und überzeugend an den Texttyp angepasst. In den Aufgaben 1 und 3 können die Einleitungsphrasen und Grußformel verbessert werden“.

6.3.5 Aspekte zum Gesamteindruck, zum Textfluss, zu kommunikative Strategien und zu Sonstiges

Die Hauptkategorien *Gesamteindruck*, *Textfluss*, *kommunikative Strategien* sowie *Sonstiges* enthalten keine weiteren Subkategorien und werden hier gemeinsam betrachtet. Aspekte, die zu diesen Kategorien gehören, werden kaum von den GER-Bewertenden kommentiert, nur von den schwedischen Bewertenden. Die Analyse erbrachte folgende Befunde bezüglich der schwedischen Bewertenden:

Tab. 27: *Verteilung der Bewerterkommentare der schwedischen Bewertenden (N = 180) auf die Kategorien Gesamteindruck, kommunikative Strategien, Textfluss und Sonstiges (Anzahl der Segmente)*

Kategorien	*negativ*	*gemischt*	*positiv*	***Gesamt***
Gesamteindruck	24	20	67	**111**
kommunikative Strategien	1	2	12	**15**
Textfluss	6	2	29	**37**
Sonstiges	2	6	11	**19**
Gesamt	**33**	**30**	**119**	**182**

Wie aus Tab. 27 hervorgeht, kommentieren die schwedischen Bewertenden relativ oft einen *Gesamteindruck*. Eine deutlich geringere Anzahl von Segmenten der schwedischen Bewertenden können der Kategorie *Textfluss,* d. h. die Kompetenz, sich mühelos und natürlich auszudrücken, zugeordnet werden. Noch weniger Kommentare können abschließend in die Kategorien *Sonstiges* und *kommunikative Strategien* eingeordnet werden.

Die große Mehrheit der Kommentare zum *Gesamteindruck* sind positiv, wobei jeweils ein Viertel als negativ oder gemischt zu betrachten ist. In einem Gesamteindruck ist es folglich häufiger, einen positiven Ton anzuschlagen. Aus Tab. 27 wird ferner ersichtlich, dass auch die Kommentare sämtlicher übriger Kategorien hauptsächlich positiv sind, aber sowohl negativ als auch gemischt eingeschätzte Kommentare sind im untersuchten Material vorhanden. Dies deutet darauf hin, dass diese Aspekte, die in den Urteilen der schwedischen Bewertenden nicht so oft vorkommen, eher dann berücksichtigt werden, wenn sie in den Schülerleistungen positiv zum Vorschein kommen oder deutlich den Text verbessern. Dies steht im Kontrast zu anderen Aspekten, wie z. B. der *Textlänge* und der *Verständlichkeit*.

Hierbei lässt sich dementsprechend ein deutlicher Unterschied zwischen den Bewertergruppen beobachten: Bis auf eine Ausnahme in der Kategorie *Sonstiges* können, wie bereits erwähnt, keine Segmente der GER-Bewertenden in diese Hauptkategorien eingeordnet werden. Das einzige Beispiel zur Kategorie *Sonstiges* wird in der qualitativen Analyse wieder aufgegriffen. Die meisten hier erwähnten Aspekte sind auch nicht im Bewertungsraster der GER-Bewertenden vertreten, wobei jedoch Aspekte zu *kommunikativen Strategien* und zur *Flüssigkeit*, vor allem im Hinblick auf die mündliche Sprachkompetenz, explizit im GER beschrieben sind (vgl. Europarat 2001: 63–69; 129).

Gesamteindruck

Die schwedischen Bewertenden formulieren relativ oft in ihren Urteilen einen ***Gesamteindruck*** der Lernproduktionen. Diese Kommentare beziehen sich auf einen globalen Eindruck des gesamten Textes (vgl. Beispiele 6.68 und 6.69) bzw. der einzelnen Teilaufgaben (vgl. Beispiel 6.70). Auffallend viele dieser Kommentare auf globaler Ebene gelten der allgemeinen kommunikativen Qualität in den Textproduktionen (vgl. Beispiele 6.71 und 6.72):

(6.68) sehr gute Lösungen. (Slps16-4-A, Lehrkraft)
(6.69) Det du har skrivit fungerar ganska bra. Lite trassel på slutet.[133] (Cllu4-3-F, Lehrkraft)
(6.70) Del 2: Mindre bra del. […] Del 3: Mycket bra [och redigt skriven formell del].[134] (Gphs5-3-B, ext. schwed. Bewert. 1)
(6.71) kommunikationen löper på bra.[135] (Vmeg5-3-D, ext. schwed. Bewert. 2)
(6.72) Eleven […] får fram det som ska sägas. […] Del 3: ej kommunicerande.[136] (Kiiu2-4-E, ext. schwed. Bewert. 1)

Übergreifende Kommentare zum Gesamteindruck sind oft relativ vage, z. B. „ok“, „gut“ oder „sehr gut“. Wie an den letzteren Beispielen ersichtlich, scheint aber auch die handlungsorientierte Perspektive in den Vordergrund zu treten, indem die kommunikativen Fertigkeiten von den schwedischen Bewertenden in ihren Kommentaren häufig auf globaler Ebene hervorgehoben werden.

---

133 „Was du geschrieben hast, funktionierst ganz gut. Einige Schwierigkeiten am Ende“.
134 „Teil 2: Der Teil ist weniger gut. […] Teil 3: Sehr guter und klar geschriebener formeller Teil“.
135 „Die Kommunikation läuft gut“.
136 „Der Schüler/die Schülerin […] bringt das, was er/sie sagen will, hervor. […] Teil 3: nicht kommunikativ“.

Darüber hinaus enthalten einige Kommentare dieser Kategorie Informationen über das Erreichen bzw. Nicht-Erreichen einer spezifischen Note (vgl. Beispiel 6.73) sowie die Stärken oder Schwächen einzelner Prüfungsteile (vgl. Beispiel 6.74):

(6.73) Räcker ej till för att nå godkänd nivå![137] (Imns4-3-F, ext. schwed. Bewert. 1)
(6.74) Uppgift 1 är utförligt skriven.[138] (Kefu5-4-E, Lehrkraft)

Diese Kommentare beziehen sich auf einen globalen Eindruck der Leistungen und nehmen zugleich Bezug auf die Benotung. In diesen Fällen können die Kommentare daher eher als Metakommentare zu den Überlegungen der Bewertenden zur der abschließenden Notengebung verstanden werden.

Kommunikative Strategien

Für die schwedischen Bewertenden ist die Kategorie ***kommunikative Strategien*** der in der vorliegenden Arbeit am wenigsten beachtete Aspekt. Da dieser Aspekt in Urteilen explizit erwähnt wird und sich nicht in die übrigen Kategorien einordnen lässt, sind Kommentare des Aspekts in einer eigenen Kategorie verblieben. Kommentare zu *kommunikativen Strategien* beziehen sich auf generalisierende Aussagen über die Verwendung kommunikativer Strategien in den Textproduktionen (vgl. Beispiele 6.75 und 6.76). In geringerem Maße kommentieren die Bewertenden kommunikative Strategien im Hinblick auf einzelne Phänomene, wie z. B. den Wortschatz (vgl. Beispiel 6.77):

(6.75) Väljer och använder i huvudsak fungerande strategier som i viss mån löser problem och förbättrar interaktionen.[139] (Crmu17-4-F, Lehrkraft)
(6.76) Vågar & utmanar – ibland funkar det, ibland inte.[140] (Smvl2-5-B, ext. schwed. Bewert. 2)
(6.77) Hittar kreativa lösningar när orden saknas (umwechslungsreich – abwechslungsreich).[141] (Kasv3-5-C, ext. schwed. Bewert. 2)

Die Mehrheit dieser Kommentare sind, wie oben ersichtlich, positive Einschätzungen. Dies kann darauf hindeuten, dass die Verwendung *kommunikativer*

137 „Reicht nicht für ein ausreichendes Niveau aus!“.
138 „Aufgabe 1 ist ausführlich geschrieben“.
139 „Wählt und verwendet hauptsächlich funktionierende Strategien, die zum Teil Probleme lösen und die Interaktion verbessern“.
140 „Wagt etwas & stellt sich Herausforderungen – manchmal gelingt das, manchmal nicht“.
141 „Findet kreative Lösungen, wenn die Wörter fehlen (umwechslungsreich – abwechslungsreich)“.

Strategien dann berücksichtigt wird, wenn diese Verwendung deutlich im Schülertext zum Vorschein kommt. Dies scheint allerdings nicht immer der Fall zu sein: einige Kommentare zu *kommunikativen Strategien* stammen aus analytischen Bewertungsrastern, die von einzelnen schwedischen praktizierenden Lehrkräfte bei der Bewertung verwendet wurden (vgl. Beispiel 6.75). Diese analytischen Bewertungsraster basieren auf den schwedischen Wissensanforderungen, zu welchen *kommunikative Strategien* gehören (vgl. Kap. 2.2.3)· Auch wenn die Verwendung *kommunikativer Strategien* somit in den schwedischen Bewertungskriterien dargestellt wird, ist nicht immer klar, wie sich dieser Aspekt in den Textproduktionen bewerten lässt. *Kommunikativen Strategien* sind aber von den schwedischen Bewertenden nicht häufig berücksichtigt und Aspekte zu Strategien werden, wenn sie in einem vorgegebenen Bewertungsraster vorkommen, ohne konkreten Texthinweis im Raster markiert.

Textfluss

Einige wenige Kommentare richten ihre Aufmerksamkeit auf den ***Textfluss*** in den Textproduktionen. Die meisten der vorzufindenden Befunde zum *Textfluss* beziehen sich auf einen globalen Eindruck der Flüssigkeit in den gesamten Textproduktionen (vgl. Beispiele 6.78 und 6.79) oder in den einzelnen Teilaufgaben (vgl. Beispiel 6.80). Die Bewertenden kommentieren gelegentlich auch den Einfluss anderer Aspekte auf den Textfluss, wie sprachliche Schwierigkeiten (vgl. Beispiel 6.81):

(6.78) Texterna har ett gott flyt.[142] (Kbtu25-4-A, ext. schwed. Bewert. 2)
(6.79) *Med flyt*: Texten har ett naturligt flöde som gör att läsaren kan följa den i princip obehindrat.[143] (Kasv3-5-A, Lehrkraft)
(6.80) Aufgabe 1: Ej flyt.[144] (Sons4-3-F, ext. schwed. Bewert. 1)
(6.81) [Språkliga brister] gör att texten saknar flyt i tillräcklig grad.[145] (Rjrv2-5-F, ext. schwed. Bewert. 2)

Die Mehrheit der Kommentare zum Textfluss sind dennoch Kommentare auf globaler Ebene und diese werden häufig nicht weiter beleuchtet oder in Verbindung mit anderen Aspekten gesetzt (z. B. Beispiel 6.78). Zu dieser Kategorie gehören auch aus einem vorgegebenen Bewertungsraster stammende

142 „Guter Textfluss in den Texten".
143 „Flüssig: Der Text hat einen natürlichen Textfluss, der es dem Leser ermöglicht, ihm im Prinzip problemlos zu folgen".
144 „Ohne Textfluss".
145 „Sprachliche Mängel führen zu einem nicht ausreichenden Textfluss".

Kommentare (vgl. Beispiel 6.79). Wenn der Textfluss in Verbindung mit anderen Phänomenen im Text gesetzt wird, sind diese Kommentare etwas häufiger negativ formuliert (vgl. Beispiel 6.81). Obwohl Aspekte des Textflusses gemäß den schwedischen Wissensanforderungen (vgl. Kap. 2.2.3) nur auf höheren Notenstufen und Niveaus vorkommen, sind Kommentare zum Textfluss auch auf niedrigen Niveaus zu finden (vgl. Beispiel 6.80).

Sonstiges

Insgesamt gehören nur sehr wenige Kommentare zur Hauptkategorie ***Sonstiges***. Die Ergebnisse der Analyse zeigen eine Diversität, die keinen einheitlichen Trend erkennen lässt. Es handelt sich in den Kommentaren u. a. um verschiedene Ausdrucksweisen (vgl. Beispiele 6.82–6.84):

(6.82) En del kreativa påhitt i brevet.[146] (Vnjg2-3-C, ext. schwed. Bewert. 1)
(6.83) Eleven uttrycker sig modigt.[147] (Hjbt-3-E, Lehrkraft)
(6.84) Humoristisk, [men icke kommunikativ].[148] (Shfg3-4-F, ext. schwed. Bewert. 1)

Die Bewertenden kommentieren hierbei Ausdrücke von Kreativität, Mut oder Humor in den Textproduktionen. Kommentare dieser Art kommen jedoch relativ selten vor und sind ausschließlich in den Bewerterurteilen der schwedischen Bewertenden zu finden.

Zur Kategorie *Sonstiges* gehören zudem Kommentare, die aus einer analytischen Herangehensweise von einer der schwedischen Lehrkräfte stammen. Hierbei wurde ein analytisches Bewertungsraster verwendet, wobei auch die Fähigkeit der Lernenden, begründete Verbesserungen an den Texten vorzunehmen, ausgewertet wurde (vgl. Beispiel 6.85):

(6.85) *Välgrundade förbättringar:* Eleven bearbetar och gör välgrundade förbättringar av egna texter.[149] (Kasv3-5-A, Lehrkraft)

Kommentare hinsichtlich Verbesserungen und Bearbeitungen in den Texten (vgl. Beispiel 6.85) gehören zwar zu den schwedischen Wissensanforderungen im Lehrplan für die zweite Fremdsprache (vgl. Kap. 2.2.3), inwieweit die Lernenden ihre Textproduktionen bearbeitet und verbesset haben, lässt sich aber schwer in diesem Kontext nachweisen und beurteilen. In welchem Umfang die

---

146 „Einige kreative Einfälle im Brief".
147 „Der Schüler/die Schülerin drückt sich mutig aus".
148 „Humoristisch, aber nicht kommunikativ".
149 „Wohlbegründete Verbesserungen: Der Schüler/die Schülerin verarbeitet und verbessert gut begründet seine eigenen Texte".

Schülerinnen und Schüler relevante Verbesserungen machen, kann vor allem evaluiert werden, nachdem die Lernenden ihre Texte nach einer Rückmeldung bearbeitet haben, d. h. nach der Arbeit einer formativen Bewertung. Dies ist aber hier nicht der Fall. Hierbei könnte es sich hingegen um ein Ausfüllen hinsichtlich der jeweiligen Benotung im Bewertungsraster handeln, das ebenso wie bei den *kommunikativen Strategien* reflexartig abläuft.

Des Weiteren kommen in den Urteilen zudem Metakommentare der Bewertenden über den eigenen Beurteilungsprozess vor. Hierzu gehören Kommentare über eine Unschlüssigkeit zwischen zwei Noten und darüber, dass eine Leistung schwer zu beurteilen ist (vgl. Beispiele 6.86 und 6.87). Vergleiche mit Aufgaben aus dem bisherigen Unterricht bei einzelnen Lernenden sind zudem in einigen Kommentaren der Lehrkräfte zu finden (vgl. Beispiel 6.88):

(6.86) ser litet ut som ordblindhet / stavningssvårigheter ibland. Det får nog bli ett C som står som min bedömning. Därav min tvekan mellan D och C, tror jag.[150] (Crpu19-4-C, ext. schwed. Bewert. 1)

(6.87) Svårbedömd[151] (Srrs2-3-E, ext. schwed. Bewert. 2)

(6.88) Eleven har inte godkänt på övrigt material heller.[152] (Hobt4-3-F, Lehrkraft)

Diese Metakommentare zum eigenen Bewertungsprozess kommen ausschließlich in den Kommentaren der schwedischen Bewertenden vor. Wie an den Beispielen 6.86 und 6.87 ersichtlich, geben Bewertende manchmal ergänzende Erklärungen zur Bewertung in ihren Kommentaren. Es handelt sich dabei u. a. um Spuren von Lese-Rechtschreib-Schwächen oder um den Einfluss der Benotung bisheriger Leistungen. Eine Orientierung der Lehrkräfte an bisherigen Bewertungen (z. B. Beispiel 6.88), ein sog. *Korrektureffekt*, könnte jedoch bedeuten, dass der Bewertende positive bzw. negative Veränderungen übersieht und dass die zu bewertende Leistung nicht angemessen beurteilt wird. Kommentare dieser Art können ein Hinweis auf Korrektureffekte sein und diese könnten eine Erklärung für eine eventuelle Variabilität zwischen Bewertungen eigener Lehrkräfte und einer externen Bewertung geben. Kommentare über bisherige Leistungen im Laufe des Jahres sind jedoch selten in den Bewerterurteilen der schwedischen Lehrkräfte zu finden.

---

150 „Sieht gelegentlich ein bisschen wie Lese-Rechtschreib-Schwäche/Legasthenie aus. Es muss wahrscheinlich die Note C sein, die als meine Bewertung steht. Daher mein Zögern zwischen den Noten D und C, glaube ich".

151 „Schwer zu beurteilen".

152 „Auch die übrigen Leistungen des Schülers sind mit ‚nicht ausreichend' benotet".

Zu dieser Kategorie gehören auch Kommentare zur Aufgabenstellung. In den Urteilen wurde u. a. kommentiert, wenn die Lernenden eine Phrase aus der Aufgabe übernommen haben (vgl. Beispiel 6.89). In diesem Zusammenhang wird von den schwedischen Bewertenden zudem erwähnt, dass die Prüfungsteilnehmenden eine der Aufgaben des Tests im Vergleich zu den anderen nicht immer so gut bewältigen können (vgl. Beispiele 6.90 und 6.91):

(6.89) *direkt aus der Aufgabe (Saig6-4, GER-Bewert. 1)
(6.90) Del 2: (svårare ämne att skriva om …).[153] (Gphs5-3-B, ext. schwed. Bewert. 2)
(6.91) men detta beror delvis på att uppgiften kräver ett något mer avancerat språk.[154] (Kcku15-4-C, ext. schwed. Bewert. 2)

Die Tatsache, dass Lernende manchmal Formulierungen aus der Aufgabe übertragen (vgl. Beispiel 6.89), scheint dementsprechend bei der Bewertung eine gewisse Bedeutung zu haben. Aus den Kommentaren geht zudem hervor, dass das inhaltliche Thema sowie die sprachlichen Anforderungen der zweiten Aufgabe für die Probanden schwieriger zu lösen scheint als die erste und dritte Aufgabe der schriftlichen Prüfung (vgl. Anhang 9). Dies zeigt sich auch bei den Kommentaren zur Aufgabenerfüllung: die zweite Aufgabe wird in höherem Ausmaß als die erste und dritte Aufgabe von den Lernenden ohne eine Antwort gelassen.

6.4 Fazit

Die erste Fragestellung fokussiert auf die Konstruktkonzeptualisierung der Bewertenden und befasst sich mit der Frage, inwieweit Bewertende ähnliche oder unterschiedliche Aspekte als besonders relevant für die Beurteilung ansehen. Die vorliegenden quantitativen und qualitativen Analysen haben gezeigt, dass Bewertende in ihren Begründungen für die Benotung schriftlicher Lernproduktionen Unterschiede im Hinblick darauf aufweisen, *welche* Aspekte sie bei der Bewertung berücksichtigen, was Fragen nach der Validität aufkommen lässt. Insgesamt scheinen vor allem Aspekte der Bewertungsdimensionen *Angemessenheit*, *formale Strukturen*, *Wortschatz*, *Aufgabenerfüllung* und *Verständlichkeit* von den teilnehmenden Bewertenden berücksichtigt zu werden. Weniger Aufmerksamkeit wird dahingegen auf *kommunikative Strategien*,

153 „Teil 2: (schwierigeres Thema zu schreiben)".
154 „Aber dies hängt teilweise damit zusammen, dass die Aufgabe eine etwas fortgeschrittenere Sprache erfordert".

Textfluss, *Gesamteindruck* und eine *pauschale Beurteilung der Sprache* gerichtet. Kommentare, die als *Sonstiges* einzustufen sind, kommen selten vor.

Die Bewertergruppen, d. h. die schwedischen Bewertenden bzw. die GER-Bewertenden, unterscheiden sich aber im Hinblick darauf, zu welchem Grad sie die jeweiligen Aspekte bei der Bewertung berücksichtigen. Die Rangordnung der meistbeachteten Aspekte bei den schwedischen Bewertenden bzw. den GER-Bewertenden wird in Tab. 28 im Vergleich dargestellt:

Tab. 28: *Reihung der meistbeachteten Aspekte in den jeweiligen Bewerterurteilen der schwedischen Bewertenden (N = 180) bzw. der GER-Bewertenden (N = 120)*

Bewertergruppe	*schwedische Bewertende*	*GER-Bewertende*
Bewertungsdimensionen	Formale Strukturen	Angemessenheit
	Wortschatz	Wortschatz
	Angemessenheit	formale Strukturen
	pauschale Beurt. – Sprache	Aufgabenerfüllung
	Aufgabenerfüllung	Verständlichkeit
	Verständlichkeit	
	Gesamteindruck	

Aus diesem Vergleich ist festzustellen, dass der meistbeachtete Aspekt der GER-Bewertenden, die *Angemessenheit*, nur die drittgewöhnlichste Dimension unter den schwedischen Bewertenden ist. Darüber hinaus kann wahrgenommen werden, dass Aspekte der linguistischen Kompetenz, d. h. *formale Strukturen*, *Wortschatz* und *pauschale Beurteilung der Sprache*, für die schwedischen Bewertenden bei der Bewertung eine große Rolle zu spielen scheinen. In diesem Zusammenhang kann zudem wahrgenommen werden, dass Aspekte zur *Korrektheit und Präzision* für sowohl die schwedischen Bewertenden als auch die GER-Bewertenden bei der Beurteilung eine größere Bedeutung zu haben scheinen als Aspekte zum *Spektrum*. Darüber hinaus weisen die Ergebnisse darauf hin, dass Aspekte zur *Aufgabenerfüllung* in den Textproduktionen in größerem Ausmaß von den GER-Bewertenden beachtet werden, während ein Kommentar zum *Gesamteindruck* häufiger von den schwedischen Bewertenden abgegeben wird. Die Bewertergruppen unterschieden sich dementsprechend im Hinblick darauf, *welche* Aspekte sie kommentieren. Unterschiede können auch im Hinblick darauf gefunden werden, inwiefern die Bewertenden Bewertungsdimensionen miteinander kombinieren, z. B. bei der Dimension *Verständlichkeit*. Auch wenn beide Bewertergruppen Aspekte zur *Verständlichkeit* berücksichtigen,

setzen die GER-Bewertenden diese häufiger in Verbindung mit der Bewältigung formaler Strukturen oder mit der Wortschatzkenntnis.

Die Bewertergruppen unterschieden sich jedoch nicht nur im Hinblick darauf, *welche* Aspekte sie bei der Bewertung berücksichtigen und *wie oft* sie diese beachten. Des Weiteren zeigen sich Unterschiede der Bewertergruppen hinsichtlich einer negativen oder positiven Einschätzung der berücksichtigten Aspekte. Die GER-Bewertenden beschreiben die Aspekte in den Textproduktionen generell in höherem Grad in positiven oder gemischten Worten, während schwedische Bewertende häufiger eine negative Einschätzung der beachteten Aspekte geben. Aus den Analysen wird zudem deutlich, dass auch Unterschiede zwischen der Gruppe der schwedischen Lehrkräfte und den beiden externen schwedischen Bewertenden zu erkennen ist. Diese Unterschiede sind generell im Bereich sprachlicher Korrektheit und inhaltlicher Erfüllung der Aufgabe zu finden. Auffallend bei dieser Analyse ist vor allem, dass die Bewertungen der praktizierenden Lehrkräfte offensichtlich mehr Wert auf sprachliche Korrekturen in den Bereichen *formale Strukturen* und *Wortschatz* legen und dabei anderen Aspekten in den Textproduktionen, wie die *Aufgabenerfüllung*, generell weniger Bedeutung schenken.

7. Analyse der Bewerterübereinstimmung

Dieses Kapitel behandelt die Bewerterübereinstimmung der schwedischen Bewertenden in Bezug auf die schriftliche Sprachfähigkeit der Lernenden und damit die zweite Fragestellung der vorliegenden Arbeit: *Wie unterscheiden sich Bewertungen bezüglich der Bewerterübereinstimmung unter den schwedischen Bewertenden?* Der Fokus liegt hierbei auf der Bewertung von Lernproduktionen durch die schwedischen Bewertenden. Wie übereinstimmig sind die Bewertenden in Bezug auf die Schreibkompetenzen der schwedischen Schülerinnen und Schüler? Und nicht zuletzt: gibt es Unterschiede im Hinblick auf die Bewerterübereinstimmung zwischen praktizierenden Lehrkräften und einer externen Bewertung? Im Kapitel wird die Aufmerksamkeit auf unterschiedliche Dimensionen der Bewerterübereinstimmung, wie den *Konsens* und die *Konsistenz* bei einer Bewertung, gerichtet.

In diesem Kapitel wird die deskriptive Statistik der Bewertungen durch die schwedischen Bewertenden, d. h. durch die Gruppe der Lehrkräfte bzw. die beiden externen Bewertenden aufgeführt (Kap. 7.1). Darauf folgen Ermittlungen zur Bestimmung der Bewerterübereinstimmung bei den Bewertungen, die von den schwedischen Bewertenden abgegeben wurden (Kap. 7.2). Hierfür werden gängige Konsens- und Konsistenzmaße (vgl. Kap. 5.3.3) verwendet, um die Interraterreliabilität der Bewerterpaare zu berechnen. Um Tendenzen zur Strenge, Mitte bzw. Milde näher aufklären zu können, sind in einem ersten Schritt die Bewertungen in Kreuztabellen wiedergegeben, wobei Überstimmungen und Nichtübereinstimmungen zwischen der Gruppe der schwedischen Lehrkräfte und den jeweiligen externen schwedischen Bewertenden veranschaulicht werden. Des Weiteren werden in einem zweiten Schritt mittels einer Multifacetten-Rasch-Analyse Unterschiede bei der Bewerterstrenge zwischen den schwedischen Bewertenden vergleichsweise untersucht (Kap. 7.3). Zunächst wird ein Vergleich von Bewerterurteilen der schwedischen Bewertenden, bei denen die gleichen Schülerleistungen unterschiedlich benotet wurden, vorgestellt (Kap. 7.4). Abschließend werden die Ergebnisse kurz zusammengefasst (Kap. 7.5).

7.1 Deskriptive Statistik der Bewertungen im schwedischen Subkorpus

Zur Beantwortung der zweiten Forschungsfrage, inwiefern die Bewertungen unter den schwedischen Bewertenden konsistent erscheinen, wurde jeder der 60 Texte in einem ersten Schritt von der an der jeweiligen Gymnasialschule praktizierenden Deutschlehrkraft und in einem zweiten Schritt von zwei unabhängigen Bewertenden beurteilt. Für die Ermittlungen der Bewerterübereinstimmung werden die schwedischen Gymnasiallehrkräfte als eine einheitliche Gruppe behandelt, auch wenn diese Gruppe aus insgesamt achtzehn Individuen besteht. Dabei finden sich einige Lehrkräfte in größerem bzw. kleinerem Materialumfang wieder als andere. Die Ergebnisse können aber Hinweise auf mögliche Tendenzen geben, die noch näher untersucht werden müssen.

Das Ergebnis der Bewertungen wird gemäß den schwedischen Bewertungskriterien auf einer sechsgradigen Skala mit den Noten F bis A dargestellt. Diese Schulnoten können im schwedischen Bildungssystem in Zahlen umgewandelt werden und um die Bewertungen vergleichen zu können, wurden zunächst die Noten auf eine Skala von 1 bis 6 transformiert. Die Noten berechnen sich wie folgt: Die Note F entspricht einem Punkt und die Note A sechs Punkten, die dazwischenliegenden Noten E, D, C und B entsprechen zwei, drei, vier bzw. fünf Punkten. In Tab. 29 sind die Mittelwerte bzw. die Standardabweichungen der Bewertungen aus der Gruppe der schwedischen Lehrkräfte ($N = 60$) und von den beiden externen schwedischen Bewertenden (jeweils $N = 60$) dargestellt:

Tab. 29: *Deskriptive Statistik (Mittelwerte und Standardabweichungen) hinsichtlich der schwedischen Bewertungen nach Fremdsprachenstufen (N = 180)*

Sprachstufe	*Gruppe der Lehrkräfte*		*ext. schwed. Bewert. 1*		*ext. schwed. Bewert. 2*	
	M	Std	M	Std	M	Std
Tyska 3	3,20	1,91	2,60	1,88	2,50	1,57
Tyska 4	3,25	1,97	2,50	1,43	2,40	1,68
Tyska 5	3,40	1,90	2,50	1,57	2,80	1,51
Gesamt	**3,28**	**1,90**	**2,53**	**1,61**	**2,57**	**1,57**

Tab. 29 zeigt, dass sich die Mittelwerte pro Stufe zwischen den Werten 2,4 und 3,4 bewegen. Zwischen den unterschiedlichen Fremdsprachenstufen können generell keine größeren Unterschiede wahrgenommen werden: Der Mittelwert für *Tyska 5* ist zwar im Vergleich zu den anderen beiden Fremdsprachenstufen bei der Gruppe der Lehrkräfte und beim zweiten externen

Bewertenden etwas höher. Wenn allerdings beachtet wird, dass vor allem eher motivierte Schülerinnen und Schüler den Kurs *Tyska 5* besuchen, ist der höhere Mittelwert dieser Stufe nicht auffällig hoch. Die Mittelwerte zwischen der Gruppe von Lehrkräften und den beiden externen Bewertenden unterscheiden sich aber deutlich. Die Werte verteilen sich wie folgt: In der Gruppe der Lehrkräfte zeigte sich insgesamt ein Gesamtmittelwert von 3,28, bei der/dem ersten externen Bewertenden zeigte sich ein Mittelwert von 2,53 und beim zweiten externen Bewertenden zeigte sich ein Mittelwert von 2,57. Der Vergleich der Gesamtmittelwerte zeigt dementsprechend eindeutige Unterschiede zwischen den Beurteilungen der Gruppe von praktizierenden Lehrkräften einerseits und denen der externen schwedischen Bewertenden andererseits. Die Gruppe der Lehrkräfte hat die Schülertexte eine halbe bis fast eine ganze Notenstufe höher benotet als die externe Bewertung, was auf Tendenzen zur Milde bzw. Strenge und zum Teil unterschiedliche Bewerterprofile hindeuten kann.

Die in Tab. 29 zusammengefassten Standardabweichungen weisen für die Gruppe von Lehrkräften eine leichte Tendenz zu höheren Werten auf, was darauf hindeuten könnte, dass die Bewertungen dieser Gruppe auf der Notenskala anders verteilt sind als die Bewertungen der externen Bewertenden. Zu beachten ist hierbei die Tatsache, dass die Bewertungen der Gruppe der Lehrkräfte in der vorliegenden Arbeit aufgrund einer bewussten Auswahl, überwiegend mit den Noten A, C, E und F, spezifisch selektiert sind, damit Textproduktionen mit divergierenden Noteneinstufungen im Material vertreten sind.

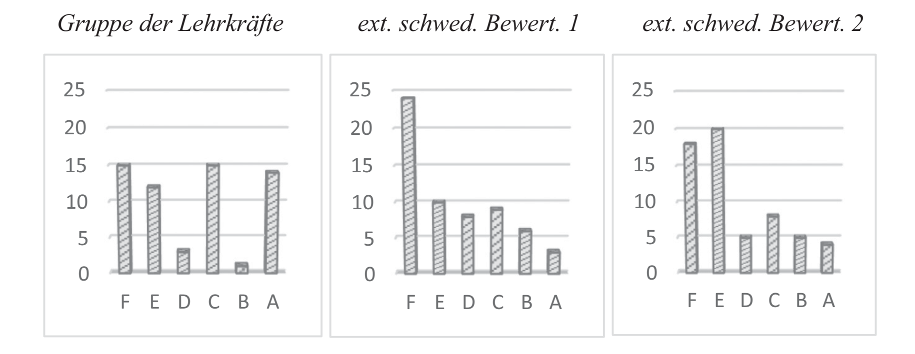


Abb. 9: Verteilung der Bewertungen über die Notenstufen (F–A) durch die Gruppe der Lehrkräfte ($N = 60$) und die zwei externen Bewertenden (jeweils $N = 60$)

Um die Distribution der Notengebung zu veranschaulichen, ist die Verteilung der Bewertungen über die sechs Notenstufen pro Bewertendem, d. h.

durch die Gruppe der Lehrkräfte bzw. die beiden externen Bewertenden, in Abb. 9 wiedergeben (Anzahl der Noten F–A):

Bei der Verteilung der Noten kann festgestellt werden, dass sich die Bewertungen der externen Bewertenden auf sämtliche Noten verteilen und sie damit die ganze Notenskala verwenden. Allerdings wird deutlich, dass die externen Bewertenden die Schülertexte im Vergleich zur Gruppe der Deutschlehrkräfte niedriger bewerten. Hierbei dominiert bei der/dem ersten externen Bewertenden eindeutig die Vergabe der Note F (*Tyska 3*: 10/20, *Tyska 4:* 7/20 und *Tyska 5:* 7/20), während die noch bestandene Note E beim zweiten externen Bewertenden, gefolgt von der Note F, überwiegt. Die Textproduktionen werden wiederum von den externen Bewertenden seltener auf die Note A eingestuft: Während insgesamt vierzehn Schülertexte von der Gruppe der Lehrkräfte die höchste Note A erhalten haben, wurden von den externen Bewertenden lediglich drei bzw. vier Texte die höchste Note zugeteilt. Dies kann als eine Neigung zur Extremtendenz, in diesem Fall aber nur in Richtung Noten im niedrigeren Bereich, verstanden werden. Die/der zweite unabhängige Bewertende zeigt zudem bei der restlichen Benotung eine leichte Zentraltendenz, d. h. die Neigung, die mittleren Notenstufen einer mehrstufigen Skala zu vergeben. Hierbei wird die mittlere Note C häufiger vergeben als die anderen mittleren Noten D und B. Die/der zweite externe Bewertende scheint insgesamt gerade bei der Vergabe der Note E nicht so streng wie der erste externe Bewertende zu sein.

7.2 Konsens und Konsistenz schwedischer Bewertender

Für die Berechnungen der Bewerterübereinstimmung zwischen der Gruppe der Lehrkräfte und den jeweiligen schwedischen Bewertenden wurden gängige Reliabilitätsmaße verwendet, für die drei Bewerterpaare jeweils drei Konsens- und Konsistenzkoeffizienten. Neben der prozentualen Übereinstimmung (PÜ) sind für die Konsensschätzungen auch die Konsensmaße Cohens Kappa und Cohens gewichtetes Kappa, zwei Maße zufallskorrigierter Übereinstimmung, zwischen je zwei Bewertenden berechnet worden. Zur Schätzungen der Konsistenz wurden hierzu die Koeffizienten für die Korrelationsmaße Spearman's Rho und Kendalls Tau-b sowie für Cronbachs Alpha, ein Maß für die interne Konsistenz, ermittelt.

In Tab. 30 sind die Konsens- sowie die Konsistenzwerte für die Bewertungen der schwedischen Bewertenden paarweise aufgeführt (Paar 1: Gruppe der Lehrkräfte – ext. schwed. Bewertender 1; Paar 2: Gruppe der Lehrkräfte – ext. schwed. Bewertender 2; Paar 3: ext. schwed. Bewertender 1 – ext. schwed. Bewertender 2):

Tab. 30: *Ergebnisse für Konsens- und Konsistenzmaße der schwedischen Bewertenden*

Paare	N	PÜ	Cohens Kappa	Gewichtetes Kappa	Spearmans Rho	Kendalls Taub	Cronbachs Alpha
1	60	38 %	.24	.55	.84	.75	.90
2	60	37 %	.23	.54	.83	.74	.89
3	60	60 %	.49	.73	.90	.82	.94

In der Tabelle enthält die dritte Spalte die Ergebnisse der prozentualen Übereinstimmung. Diese reichen von 37 % bzw. 38 % bis zu 60 %, wobei aber erst eine Übereinstimmung ab 70 % als zufriedenstellend gilt (vgl. Stemler 2004). Die niedrigsten Werte zeigen die Bewerterpaare zwischen der Gruppe der Lehrkräfte und einem externen Bewertenden, während der höchste Wert zwischen den beiden externen Bewertenden zu finden ist. Ähnliche Befunde finden sich in der vierten Spalte für die Kappa-Werte. Diese liegen zwischen .23 und .49, wobei der höchste Wert .49 wieder zwischen den beiden externen Bewertenden vorliegt. Ein Kappa-Wert über .4 kann aber als akzeptabel gelten (vgl. Landis & Koch 1977), wobei die beiden anderen Kappa-Werte auf eine mangelnde Übereinstimmung hindeuten. Der gewichtete Kappa-Koeffizient zeigt ebenfalls niedrigere Werte zwischen der Gruppe der Lehrkräfte und den beiden externen Bewertenden, wobei sämtliche Werte aber als zufriedenstellend gelten könnten. Für die beiden externen Bewertenden weist der Wert sogar auf eine gute Übereinstimmung hin. Insgesamt zeigen die Konsenswerte, dass sich der Grad der Übereinstimmung zwischen der Gruppe der Lehrkräfte und den jeweiligen externen Bewertenden auf einem niedrigeren bis bestenfalls akzeptablen Niveau befindet. Zu bemerken ist allerdings, dass die Werte für das Bewerterpaar mit den beiden externen Bewertenden deutlich höher als Bewerterpaare mit der Gruppe der Lehrkräfte liegen und auf eine zufriedenstellende bis gute Übereinstimmung hinweisen.

Betrachtet man andererseits die Konsistenzwerte, kann dennoch insgesamt festgestellt werden, dass die geforderte Reliabilitätshöhe von .70 nach Barrett (2001) und Stemler (2004) erreicht ist. Die Ergebnisse für Spearman's Rho in der fünften Spalte zeigen ziemlich hohe Werte von .83 bis .90, während Kendalls Tau-b in der sechsten Spalte, dessen Werte häufig etwas niedriger ausfallen als die Spearman-Rangkorrelationen, von .74 bis zu .82 reicht. Die Berechnungen des internen Konsistenzmaßes Cronbachs Alpha in der letzten Spalte ergeben Werte zwischen .89 und .94, was auf eine hohe interne Konsistenz hindeutet.

Ähnlich wie bei den Konsensmaßen der schwedischen Bewertenden fällt auf, dass die Werte für das Bewerterpaar mit den beiden externen Bewertenden auf einem höheren Niveau liegen als für Bewerterpaare mit der Gruppe der Deutschlehrkräfte.

7.3 Schwedische Bewertende: Milde- bzw. Strengetendenzen

Um den Differenzen der Bewerterübereinstimmung nachgehen zu können, sind die Ergebnisse der Bewertungen paarweise in Kreuztabellen dargestellt. In Tab. 31 sind die Bewertungen der Gruppe der Lehrkräfte und des ersten externen Bewertenden aufgeführt:

Tab. 31: *Kreuztabelle mit Bewertungen der Textproduktionen durch die Gruppe der Lehrkräfte und die/den externen schwedischen Bewertende/n 1*

Gruppe der Lehrkräfte	*ext. schwedischer Bewertender 1*						
	F	**E**	**D**	**C**	**B**	**A**	**Gesamt**
F	15						**15**
E	6	3	2	1			**12**
D		3					**3**
C	3	4	5	2	1		**15**
B				1			**1**
A			1	5	5	3	**14**
Gesamt	**24**	**10**	**8**	**9**	**6**	**3**	**60**

In der Kreuztabelle zeigt die Zeilensumme ganz rechts die Verteilung der Noten durch die Gruppe der schwedischen Lehrkräfte. Diese verteilen sich nach der bewussten Textauswahl auf die Noten F, E, C und A (vgl. Kap. 5.2). Die Spaltensumme ganz unten in der Kreuztabelle zeigt umgekehrt die Verteilung der Noten durch die/den ersten schwedischen Bewertende/n, wohin weniger Variation zu finden ist und bei der Vergabe der Noten niedrigere über höhere Notenstufen überwiegen. Die grau unterlegten Felder diagonal in der Tabelle weisen auf die Textproduktionen hin, die pro Note die gleiche Einstufung erhalten haben. Beispielsweise haben insgesamt zwölf Textproduktionen von der Gruppe der Lehrkräfte die Note E erhalten. Daraus wurde drei Schülerleistungen von sowohl der Gruppe der Lehrkräfte als auch von der/dem ersten externen Bewertenden eine Note E zugeteilt. Ferner haben insgesamt neun Texte, denen von der Gruppe der Lehrkräfte eine Note E gegeben wurde, von der/dem ersten externen Bewertenden eine abweichende Note erhalten: sechs

Texte sind niedriger eingestuft, während drei Texte mit D bzw. C benotet wurden.

Die prozentuale Übereinstimmung (PÜ) zwischen der Gruppe der Lehrkräfte und der/dem externen Bewertenden 1 beträgt, wie auch aus Tab. 30 zu entnehmen ist, 38 % (d. h. 23 Übereinstimmungen und 37 Nichtübereinstimmungen), was ein relativ niedriger Wert ist. Aus der Kreuztabelle 31 oben wird ersichtlich, dass die/der erste externe Bewertende generell eine niedrigere Einstufung vornimmt, was jedoch die Berechnungen der Mittelwerte und die Abbildungen der Notenverteilungen oben bereits angedeutet haben (vgl. Tab. 29 bzw. Abb. 9). Die insgesamt 37 Nichtübereinstimmungen finden sich bei allen Notenstufen außer der niedrigsten Note F. Die meisten Nicht-Übereinstimmungen betreffen die Notenstufen C und A, aber auch die E-Note. Bei den Notenstufen C und A unterscheiden sich die Einstufungen manchmal sogar um drei Notenstufen in Richtung einer niedrigeren Note für die Bewertungen des externen Bewertenden. Hier ist festzustellen, dass die/der erste externe Bewertende die Schülertexte strenger als die Gruppe der Lehrkräfte bewertet hat und dass die Gruppe der Lehrkräfte somit den Leistungen häufiger eine befriedigende Note erteilt hat.

In Tab. 32 sind die Bewertungen durch die Gruppe der schwedischen Lehrkräfte erneut aufgeführt, diesmal jedoch im Vergleich zu der/dem zweiten externen Bewertenden:

Tab. 32: *Kreuztabelle mit Bewertungen der Textproduktionen durch die Gruppe der Lehrkräfte und die/den externen schwedischen Bewertende/n 2*

Gruppe der Lehrkräfte	*ext. schwedischer Bewertender 2*						
	F	**E**	**D**	**C**	**B**	**A**	**Gesamt**
F	12	3					**15**
E	5	5	2				**12**
D		3					**3**
C	1	8	4	1	1		**15**
B				1			**1**
A			1	5	4	4	**14**
Gesamt	**18**	**19**	**7**	**7**	**5**	**4**	**60**

Ein ähnliches Bild ergibt sich, wenn die Einstufungen der Gruppe der Deutschlehrkräfte denjenigen der/des zweiten externen Bewertenden in der Kreuztabelle gegenübergestellt werden. Auch hier zeigen die Spaltensummen

ganz unten in der Tabelle, dass die/der externe Bewertende in höherem Ausmaß als die Gruppe der Lehrkräfte eine niedrigere Bewertung abgibt.

Betrachtet man die prozentuale Übereinstimmung für die Gruppe der Lehrkräfte und die/den Bewertende/n 2 von 37 % (d. h. 22 Übereinstimmungen und 38 Nichtübereinstimmungen), kann festgestellt werden, dass die beiden externen Bewertenden im Vergleich zur Gruppe von Lehrkräften ähnliche Übereinstimmungsraten aufweisen. Die insgesamt 38 Nichtübereinstimmungen sind bei allen Notenstufen zu finden. Ebenfalls, wie beim Vergleich Lehrkräfte – die/der erste externe Bewertende ist die Tendenz, dass die/der zweite unabhängige Bewertende deutlich in Richtung einer niedrigeren Einstufung geht. Hierbei zeigt sich aber bei dieser/diesem externen Bewertenden eine schwache Tendenz, gelegentlich bei der Note F positiver zu bewerten. Wie bei der/dem ersten unabhängigen Bewertenden betrifft eine überwiegende Anzahl der Nichtübereinstimmungen die Notenstufen C und A, aber relativ häufig auch die Note E. Gegenüber den Einstufungen der Notenstufen C und A durch die Gruppe der Lehrkräfte bewertete die/der zweite externe Bewertende, wie auch im vorherigen Vergleich der erste, in einzelnen Fällen sogar drei Notenstufen niedriger. Es handelt sich dabei oft, aber nicht immer, um die gleichen Textproduktionen, die auch von der/dem ersten externen Bewertenden niedriger eingestuft wurden.

Abschließend sollen die Einstufungen der beiden unabhängigen externen Bewertenden veranschaulicht werden. Tab. 33 zeigt dementsprechend einen Vergleich zwischen den Bewertungen dieser beiden Bewertenden:

Tab. 33: *Kreuztabelle mit Bewertungen der Textproduktionen durch die externen schwedischen Bewertenden 1 und 2*

ext. schwed. Bewertender 1	*ext. schwedischer Bewertender 2*						
	F	**E**	**D**	**C**	**B**	**A**	**Gesamt**
F	18	6					**24**
E		9	1				**10**
D		4	3	1			**8**
C			3	3	2	1	**9**
B				2	2	2	**6**
A				1	1	1	**3**
Gesamt	**18**	**19**	**7**	**7**	**5**	**4**	**60**

Die Kreuztabelle der Einstufungen durch die beiden externen Bewertenden stellt ein anderes Bild dar als die Vergleiche mit der Gruppe der

Gymnasiallehrkräfte. Mit einer prozentualen Übereinstimmungsrate von 60 % (d. h. 36 Übereinstimmungen und 24 Nichtübereinstimmungen) ist der Wert relativ gut. Die insgesamt 24 Nichtübereinstimmungen verteilen sich auf die ganze Notenskala. Wie bereits in Abb. 9 deutlich wurde, zeigt die/der zweite Bewertende eine Tendenz, die Lernproduktionen niedrigerer Niveaus vergleichsweise milder zu bewerten: ein Drittel der F-Texte bei der/dem ersten Bewertenden hat von der/dem zweiten Bewertenden eine Note E erhalten. Unterscheide um mehr als eine Notenstufe kommen für diese beiden Bewertenden eher selten vor. Nur bei den Noten C und A unterscheiden sich die Einstufungen manchmal mehr als eine Notenstufe.

Um mögliche Milde-Strenge-Tendenzen der schwedischen Bewertenden untersuchen zu können, wurden ihre Bewertungen zudem mithilfe einer Multifacetten-Rasch-Analyse untersucht. Abb. 10 gibt einen Überblick über den Facettenraum, wodurch Vergleiche zwischen den Facetten „Prüfungsteilnehmende“, „Bewertende“, und „Benotung“ möglich sind:

```
+-------------------------------------------------------------------------------------------------------------------------+
|Measr|+Prüfungsteilnehmende                                                                           |-Bewertende|ERGEB|
|-----+------------------------------------------------------------------------------------------------+-----------+-----|
|   6 + Gols-6                                                                                          +           + (5) |
|     | Kbtu-25 Sces-17 Öhfl-3                                                                          |           |  A  |
|     |                                                                                                 |           |     |
|   5 + Gphs-5  Ihsu-1  Kasv-3                                                                          +           + --- |
|     |                                                                                                 |           |     |
|     |                                                                                                 |           |     |
|   4 + Cemu-14 Slps-16 Tect-4  hjjg-6                                                                  +           +  4  |
|     |                                                                                                 |           |  B  |
|     | Kcku-15 Smvl-2                                                                                  |           |     |
|   3 +                                                                                                 +           + --- |
|     |                                                                                                 |           |     |
|     | Crpu-19 Vnjg-2  ekls-1                                                                          |           |     |
|   2 +                                                                                                 +           +     |
|     |                                                                                                 |           |  3  |
|     | Klju-1                                                                                          |           |  C  |
|   1 +                                                                                                 + ext. 1    +     |
|     | Vmeg-5                                                                                          | ext. 2    |     |
|     |                                                                                                 |           |     |
*   0 *                                                                                                 *           * --- *
|     | Ghhs-4  Kinv-5                                                                                  |           |     |
|     |                                                                                                 |           |     |
|  -1 + Hchg-2  Sing-1  Srrs-2  Vedg-3  Vjrg-24                                                         +           +  2  |
|     |                                                                                                 |           |  D  |
|     |                                                                                                 | Lehrkraft |     |
|  -2 + Eles-2  Ilgs-11 Kiiu-2  Sess-13                                                                 +           + --- |
|     |                                                                                                 |           |     |
|     | Cswu-1  Hjlg-4  Hvbg-3  Imls-9  Kpnu-28 Twpt-3                                                  |           |     |
|  -3 +                                                                                                 +           +     |
|     |                                                                                                 |           |     |
|     |                                                                                                 |           |  1  |
|  -4 + Hjbt-5  Hmlt-2  Kefu-5  Vwbg-25                                                                 +           +  E  |
|     |                                                                                                 |           |     |
|     |                                                                                                 |           |     |
|  -5 + Geks-8                                                                                          +           +     |
|     |                                                                                                 |           |     |
|     |                                                                                                 |           | --- |
|  -6 +                                                                                                 +           +     |
|     |                                                                                                 |           |     |
|     | Imns-4  Pnmj-1  Saig-6  Slsk-1  Soal-1  Sons-4  Vmlg-21 Ösnl-9                                  |           | (0) |
|  -7 + Cllu-4  Crau-16 Hobt-4  Rdsv-1  Rjrv-2  Shfg-3  Shfg-3  Sjel-3  Splg-8  Vegg-35 Vmcg-1  Örkl-1  +           +  F  |
|-----+------------------------------------------------------------------------------------------------+-----------+-----|
|Measr|+Prüfungsteilnehmende                                                                           |-Bewertende|ERGEB|
+-------------------------------------------------------------------------------------------------------------------------+
```

Abb. 10: Ergebnisse der Multifacetten-Rasch-Analyse bei der Beurteilung fremdsprachlicher Leistungen durch die schwedischen Bewertenden ($N = 180$)

Wie in Abb. 10 ersichtlich, besteht der Facettenraum aus vier Spalten. Die erste Spalte enthält den Messwert in Form von Logits. Die höchsten Werte sind oben in der Abbildung zu finden, während sich die niedrigsten unten befinden. Die zweite Spalte gibt die Verteilung der Prüfungsteilnehmenden wieder. Hierbei sind leistungsstärkere Probanden ganz oben im positiven Bereich und weniger leistungsstarke Probanden ganz unten.

In der dritten Spalte sind die Bewertenden im Hinblick auf ihre jeweilige Bewerterstrenge wiedergegeben. Ein Wert nahe 0 zeigt, dass die Bewertenden weder streng noch mild in Relation zu den anderen Bewertenden beurteilen. Positive Logitwerte verweisen auf eine strengere Bewertung, während dahingegen negative Logitwerte auf eine Milde-Tendenz deuten. Die Gruppe der Lehrkräfte (in Abb. 10: *Lehrkraft*) unterscheidet sich mit -1.75 Logits von den externen Bewertenden (in Abb. 10 *ext. 1* bzw. *ext. 2*), die 1.00 bzw. .75 Logits aufweisen. Aus der Abbildung geht somit deutlich hervor, dass die Gruppe der Lehrkräfte milder als die externen Bewertenden beurteilt. Umgekehrt kann man auch sagen, dass die externen Bewertenden eine leichte Strenge-Tendenz aufweisen. Dieser Unterschied beträgt in etwa eine Notenstufe. Die letzte Spalte gibt das Kompetenzniveau der Prüfungsteilnehmenden, in diesem Fall in der Notenskala des schwedischen Systems, wieder. Zu bemerken ist hier, dass der Anteil von Prüfungsteilnehmenden, die von sämtlichen Bewertenden eine nicht ausreichende Note F erhalten haben, relativ hoch ist.

Bei einer Multifacetten-Rasch-Analyse sollten *Infit bzw. Outfit Mean-Square-Statistiken* (MnSq) ausgewertet werden, damit untersucht werden kann, inwiefern diese Werte zum Raschmodell passen. Diese Werte können somit über den Grad der Konsistenz der einzelnen Bewertender informieren, indem sie ermitteln, inwiefern die Bewertungen einzelner Bewertenden größere Variationen zeigen als vom Modell erwartet wird, oder nicht.

In welchem Ausmaß Beurteilungen einzelner Bewertender zu den erwarteten Beurteilungen, die ein einem bestimmten Multifacetten-Rasch-Modell generiert wurden, passen, wird von den Infit- bzw. Outfitstatistiken indiziert. Diese sollten im Intervall zwischen 0.5 und 1.5 (manchmal werden auch engere Richtwerte zwischen 0.7 und 1.3 angegeben) liegen. Die Werte der Infit- bzw. Outfitstatistiken der Multifacetten-Rasch-Analyse sind in Tab. 34 aufgeführt:

Tab. 34: *Infit- bzw. Outfitwerte der Multifacetten-Rasch-Analyse für die Bewertungen der schwedischen Bewertenden (N = 180)*

Bewertende	*N*	*Infit*	*Outfit*
Gruppe der Lehrkräfte	60	1.25	1.09
ext. schwed. Bewert. 1	60	.88	.89
ext. schwed. Bewert. 2	60	.71	.88

Die Infit- bzw- Outfitwerte im Datensatz befinden sich sämtlich im Intervall 0.5 bis 1.5 und fallen den Faustregeln folgend daher nicht ins Auge. Die Gruppe der Lehrkräfte hat mit einem Infitwert von 1.25 eine leichte Tendenz in Richtung zum *Misfit* (oder *Underfit*). Dies bedeutet, dass die Benotung der Textproduktionen 25 % mehr variiert als vom Modell erwartet wird, was wahrscheinlich aber auf die bewusste Variation der benoteten Texte bei der Textauswahl zurückzuführen ist. Die beiden externen schwedischen Bewertenden weisen mit Werten zwischen .71 und .89 dahingegen eine Tendenz zum *Overfit* auf. Diese Bewertenden zeigen bei der Benotung in etwa 20–30 % weniger Variation als erwartet auf. Die relativ große Anzahl von Texten, die mit E oder F benotet wurden, könnte eine Erklärung dafür sein (vgl. Abb. 9).

7.4 Qualitativer Vergleich von Urteilen unterschiedlicher bzw. ähnlicher Ergebnisse

Um der Frage nachgehen zu können, inwiefern Bewertende die gleichen oder unterschiedliche Aspekte bei der Bewertung einzelner Leistungen berücksichtigen und inwieweit dies als Grund für unterschiedliche bzw. ähnliche Ergebnisse angeführt werden könnte, wurden zudem vertiefende Analysen durchgeführt. Für diesen qualitativen Vergleich wurden insbesondere Bewerterurteile mit möglichst unterschiedlichen bzw. möglichst ähnlichen Ergebnissen näher betrachtet. In der folgenden Analyse wird somit untersucht, inwiefern a) Bewertungen, die geringere Übereinstimmungen aufweisen, und b) Bewertungen, bei deren Benotung die Bewertenden übereinstimmen, Differenzen und/oder Gemeinsamkeiten aufweisen. Zwei Textproduktionen, die eine besonders divergierende Benotung von den schwedischen Bewertenden erhalten haben, sind die Schülerleistungen *Hmlt2*–3 (*Tyska 3*) und *Kpnu28-5* (*Tyska 5*). Ähnliche Einstufungen haben vor allem Textproduktionen niedrigerer Niveaus, insbesondere mit einer nicht ausreichenden Note F. Da die Bewertung von Textproduktionen mittleren Niveaus im Hinblick auf die Notenvergabe weniger häufig übereinstimmt, wird im Folgenden eine der wenigen Textproduktion

mit übereinstimmender Bewertung mittleren Bereichs näher beschrieben. Es handelt sich hierbei um die Schülerleistung *Klju1-4* (*Tyska 4*), die von sämtlichen schwedischen Bewertenden eine C-Note erhalten hat. Hier folgt ein qualitativer Vergleich der Bewerterurteile dieser erwähnten Leistungen.

Bewerterurteile unterschiedlicher Bewertung

Eine Schülerleistung mit divergierendem Ergebnis zwischen den schwedischen Bewertenden im Hinblick auf die Benotung ist der Text ***Hmlt2–3: C/F/F*** (*Tyska 3*). Die Textproduktion *Hmlt2*–3 ist von der eigenen Lehrkraft mit einer Note C bewertet worden, von der/dem ersten und zweiten externen Bewertenden hat die Schülerleistung aber die Note F erhalten.[155] In ihren Begründungen der Bewertung haben sowohl die praktizierende Lehrkraft als die beiden externen schwedischen Bewertenden *Aspekte der linguistischen Kompetenz* berücksichtigt (vgl. Beispiele 7.1–7.3):

(7.1) Ordföljden är ganska bra. Bra tempus.[156] (Hmlt2-3-C, Lehrkraft)
(7.2) Dock finns vissa ok ordval & korrekta fraser.[157] (Hmlt2-3-F, ext. schwed. Bewert. 2)
(7.3) Till stora delar obegripligt språk. […] Oidiomatiskt.[158] (Hmlt2-3-F, ext. schwed. Bewert. 1)

Auch wenn sämtliche Bewertenden Aspekte der linguistischen Kompetenz beachtet haben, unterscheiden sich ihre Kommentare stark voneinander. Die Lehrkraft kommentiert *formale Strukturen* in der Textproduktion, wobei explizit die relativ gute Beherrschung von Wortstellung und Zeitformen genannt wird. Die Kommentare von der eigenen Lehrkraft sind somit deutlich positiv wertend. Die beiden Kommentare von den externen schwedischen Bewertenden sind allgemeiner formuliert und berücksichtigen zum einen Aspekte zu *Wortschatz* sowie *Korrektheit* und geben zum anderen eine *pauschale Beurteilung der Sprache*. Der erste externe Bewertende setzt hierbei die Sprache mit der *Verständlichkeit* des Textes in Verbindung und gibt in seinem negativ wertenden Kommentar an, dass die Sprache idiomatisch nicht funktioniert und in vielerlei Hinsicht unverständlich ist.

155 Beim Gesamtergebnis der GER-Bewertung hat diese Leistung insgesamt 30 Punkte erhalten und dementsprechend nicht die Anforderungen eines B1-Niveaus erfüllt.
156 „Wortfolge ist ziemlich gut. Gutes Tempus". (*Hier und im Folgenden eigene Übersetzung M.H.R.*).
157 „Es gibt jedoch gewisse akzeptable Wortwahl & korrekte Phrasen".
158 „Sprache ist in vielen Teilen unverständlich. […] Unidiomatisch".

Des Weiteren beachten sämtliche schwedische Bewertende in ihren Bewerterurteilen die *Textlänge* und wie sich diese auf die *inhaltliche Aufgabenerfüllung* und den *Gesamteindruck* bezieht (vgl. Beispiele 7.4–7.6):

(7.4) Kort text, men bra ändå.[159] (Hmlt2-3-C, Lehrkraft)
(7.5) Första uppgiften genomförs – del två och tre väl korta & utvecklar ej innehållet.[160] (Hmlt2-3-F, ext. schwed. Bewert. 2)
(7.6) För kort, ger ej den kommunikation som förväntas.[161] (Hmlt2-3-F, ext. schwed. Bewert. 1)

Die Lehrkraft ist bei der *Textlänge* der Meinung, dass die Schülerleistung trotz der wenigen Wörter gut gelungen ist. Die externen Bewertenden sehen das anders und schätzen ein, dass die Leistung aufgrund der mangelnden Wortanzahl das Erwartungsniveau für den kommunikativen *Gesamteindruck* und die *Erfüllung vom Inhalt* her nicht erreicht.

Darüber hinaus kommentiert der erste externe Bewertende die *soziokulturelle Angemessenheit* bei der letzten Aufgabe des Tests (vgl. Beispiel 7.7):

(7.7) Abschluss.[162] (Hmlt2-3-F, ext. schwed. Bewert. 1)

Hier scheint der erste externe Bewertende die Abschlussformel nicht als formell genug für eine E-Mail an den eigenen Lehrer zu betrachten. Aspekte zur *soziokulturellen Angemessenheit* werden sonst in den Bewerterurteilen der anderen beiden schwedischen Bewertenden für diesen Schülertext nicht berücksichtigt.

Zusammenfassend ist bei der Schülerleistung *Hmlt2-3* festzustellen, dass die schwedischen Bewertenden häufig in etwa die gleichen Aspekte betrachten, diese Aspekte aber zum Teil unterschiedlich gewichten, z. B. die Auswirkung der etwas kürzeren *Textlänge* oder inwiefern ein Fokus auf die *Korrektheit* oder die *Verständlichkeit* gelegt werden sollte. Aber auch die Berücksichtigung unterschiedlicher Aspekte, wie hier der *Angemessenheit*, kann wahrgenommen werden.

Divergierende Noten hat von den schwedischen Bewertenden auch die Textproduktion ***Kpnu28-5: C/E/F*** (*Tyska 5*) erhalten. Die eigene Lehrkraft hat dieser Schülerleistung die Note C gegeben, sie wurde aber von der/dem ersten und

159 „Kurzer Text, aber immer noch gut".
160 „Die erste Aufgabe wird gelöst – Die Teile zwei und drei sind sehr kurz & und realisieren den Inhalt nicht".
161 „Zu kurz, der Text gibt nicht die zu erwartende Kommunikation".
162 Der Kommentar bezieht sich auf die Abschlussformel im Text: „Tschüss!"

zweiten externen Bewertenden mit den Noten E bzw. F bewertet.[163] Sämtliche schwedischen Bewertenden beachten in ihren Bewerterurteilen die sprachliche *Korrektheit* (vgl. Beispiele 7.8–7.10):

(7.8) Eleven visar dock ganska stora grammatiska brister; t.ex. ordföljd, genus, kasus, val av hjälpverb, kongruens, viss brist inom vokabulären (t.ex. „Ich will dir sehen" och „eine Gränge"). Det mesta är begripligt och relativt välformulerat, om än väldigt enkelt uttryck, men vissa bitar är lite svåra att förstå p.g.a. brister som grammatik + vokabulär.[164] (Kpnu28-5-C, Lehrkraft)

(7.9) Vissa felval gällande vokabulär. Brister grammatiskt, men ändå relativt tydligt.[165] (Kpnu28-5-E, ext. schwed. Bewert. 2)

(7.10) Större delen på mycket inkorrekt tyska gör det hela oklart och svårt att följa vad eleven menar. Stora brister i språklig precision.[166] (Kpnu28-5-F, ext. schwed. Bewert. 1)

Die Bewertenden kommentieren wie an den Beispielen ersichtlich sprachliche Mängel in der Textproduktion und deren Einwirkung auf die *Genauigkeit* und die *Verständlichkeit* im Text. Sie unterscheiden sich aber zum Teil im Hinblick darauf, wie diese Mängel zu interpretieren sind. Die praktizierende Lehrkraft und die/der zweite externe Bewertende beachten sowohl grammatische Fehlgriffe als auch Mängel im Wortschatz. Sie kommen hierbei zum Schluss, dass der Text trotzdem relativ gut ist. Der erste externe Bewertende findet dagegen, dass die großen Mängel in Bezug auf sprachliche *Korrektheit* und *Präzision* die *Verständlichkeit* in der Schülerleistung stark beeinträchtigen.

---

163 Die Note E vom zweiten externen Bewertenden ist mit einem Minus versehen, indizierend, dass die Leistung auf der Grenze zu einer nicht ausreichenden Note liegt. Bei der GER-Bewertung wurden diesem Schülertext 75 Punkte, was ein erreichtes B1-Niveau bedeutet, zugeteilt.

164 „Der Schüler/die Schülerin weist jedoch ziemlich große grammatische Mängel auf, z. B. bei Wortfolge, Genus, Kasus, Wahl von Hilfsverben, Kongruenz und gewisse Defizite im Wortschatz (z. B. ‚Ich will dir sehen' und ‚eine Gränge'). Das Meiste ist verständlich und relativ gut formuliert, wenn auch sehr einfach formuliert, aber gewisse Teile sind wegen Mängeln wie Grammatik + Wortschatz schwer zu verstehen".

165 „Einige Fehlgriffe bezüglich des Wortschatzes. Grammatische Mängel, aber immer noch relativ deutlich".

166 „Das Meiste in sehr inkorrektem Deutsch. Dies macht der Text unklar und es wird schwierig zu verstehen, was der Schüler/die Schülerin meint. Große Mängel an sprachlicher Präzision".

Des Weiteren kommentieren zwei der schwedischen Bewertenden, in welchem Ausmaß dem/der Lernenden die *inhaltliche Aufgabenerfüllung* in der Aufgabe gelingt (vgl. Beispiele 7.11 und 7.12):

(7.11) Eleven håller sig ganska bra till ämnena.[167] (Kpnu28-5-C, Lehrkraft) meritpoänm

(7.12) Följer instruktionen även om innehållet är ngt tunt. […] Kommenterar ej fullt ut brevet.[168] (Kpnu28-5-E, ext. schwed. Bewert. 2)

Die Bewertenden kommentieren in den Beispielen, in welchem Ausmaß die Textproduktion den nachgefragten Inhalt enthält. Hierbei wird aber deutlich, dass die beiden Bewertenden die Anforderungen an den Inhalt unterschiedlich verstehen. Die Lehrkraft äußert sich hinsichtlich der Realisierung des Inhalts im Text eher positiv, während die Schülerleistung nach Ansicht des zweiten externen Bewertenden die inhaltlichen Anforderungen nicht ganz erfüllt. Der erste externe Bewertende kommentiert dagegen nicht die *inhaltliche Aufgabenerfüllung,* dagegen wird im Hinblick auf die Anforderungen der Aufgabe die *Textlänge* erwähnt. Hierbei wird die Wortanzahl als zu gering angegeben. Auf die *Textlänge* wird jedoch von den beiden anderen Bewertenden nicht eingegangen.

Ferner kommentiert lediglich die Lehrkraft die *Angemessenheit* in der Textproduktion (vgl. Beispiel 7.13):

(7.13) Eleven anpassar i viss mån språket och texten till mottagaren. […] Eleven uppfyller till relativt stor del de formella kriterierna för respektive texttyp.[169] (Kpnu28-5-C, Lehrkraft)

Hierbei werden von der praktizierenden Lehrkraft sowohl Aspekte zur *soziokulturellen Angemessenheit* als auch zur Umsetzung der *Textsorte* in positiven Worten aufgegriffen, Aspekte, die von den anderen beiden schwedischen Bewertenden nicht erwähnt werden.

An der obigen Analyse der divergierenden Benotung bei der Schülerleistung *Kpnu28-5* ist festzustellen, dass die schwedischen Bewertenden relativ häufig die gleichen Aspekte betrachten, diese aber zum Teil unterschiedlich gewichten. Dies gilt hier u. a. für die *inhaltlichen* Anforderungen und die sprachlichen

167 „Der Schüler/die Schülerin hält sich ziemlich gut an das Thema".

168 „Befolgt die Anweisungen, auch wenn der Text inhaltlich etwas dünn ist. […] Kommentiert nicht vollständig den Brief".

169 „Der Schüler/die Schülerin passt zu einem gewissen Grad die Sprache und den Text an den Empfänger an. […] Der Schüler/die Schülerin erfüllt weitgehend die formalen Kriterien für die jeweilige Textsorte".

Mängel im Hinblick auf *formale Strukturen* und *Wortschatz*. Wie in der Analyse bei der Schülerleistung *Hmlt2–3* werden Aspekte zur *Angemessenheit* nicht von allen Bewertenden berücksichtigt. Aus den Kommentaren entsteht ein Bild von einer Schülerleistung mit sowohl positiven als auch negativen Aspekten, was eventuell zur divergierenden Bewertung der jeweiligen schwedischen Bewertenden beigetragen hat.

Dies wird beispielsweise bei der Berücksichtigung von Aspekten im Hinblick auf die *Aufgabenerfüllung* deutlich: der Proband hat zwar die Anweisungen befolgt, diese aber eher knapp beantwortet. Es scheint daher für die Bewertenden schwieriger zu sein, diese Aspekte als positiv oder negativ einzuschätzen und für die Benotung zu gewichten.

Bewerterurteile ähnlicher Bewertung

In der qualitativen Analyse wurden auch Bewertungen, bei denen die Benotungen der schwedischen Bewertenden übereinstimmen, untersucht. Unter den Schülerleistungen, die von sämtlichen drei schwedischen Bewertenden mit derselben Note bewertet sind, finden sich häufig Texte mit einer nicht ausreichenden Note F.[170] Die schwedischen Bewertenden kommentieren in diesen Bewerterurteilen in der Regel sprachliche Mängel, das Fehlen einer oder zweier Teilaufgaben, mangelnde Verständlichkeit und eine eher knappe Textlänge. Zum Teil wird auch die Anpassung im Text bezüglich der Verwendung formeller oder informeller Sprache oder der Umsetzung der Textsorte kommentiert. Dieser Aspekt kommt jedoch nicht in der Mehrheit der Bewerterkommentare vor und wird zudem nicht von allen Bewertenden aufgegriffen.

Im Hinblick darauf, dass die schwedischen Bewertenden bei der Vergabe der C-Note häufig nicht übereinstimmen, ist es relevant, die Aufmerksamkeit auf das einzige Beispiel zu richten, wo die schwedischen Bewertenden bei der Notenvergabe C übereinstimmen. Die Schülerleistung ***Klju1–4: C/C/C*** (*Tyska 4*) wurde von der eigenen Lehrkraft sowie von der/dem ersten und zweiten externen Bewertenden mit der Note C bewertet.[171] Die Bewertenden beachten bei der Bewertung *Aspekte der linguistischen Kompetenz*, sowohl gute Formulierungen als auch sprachliche Mängel. Die Lehrkraft und der erste externe Bewertende

---

170 Zu bemerken ist hierbei, dass keiner der Textproduktionen, die von sämtlichen schwedischen Bewertenden mit der Note F beurteilt wurden, bei der GER-Bewertung ein erreichtes Niveau B1 erreicht hat.

171 Bei der GER-Bewertung hat diese Leistung eine Gesamtpunktzahl von 88 Punkten erhalten und somit das Niveau B1 erreicht.

berücksichtigen dabei zudem die Auswirkung sprachlicher Fehlgriffe auf die *Verständlichkeit* (vgl. Beispiele 7.14 und 7.15):

(7.14) det är dock väldigt välformulerat och begripligt […] Några grammatiska + stavfel.[172] (Klju1-4-C, Lehrkraft)

(7.15) Felaktiga meningar överlag. […] Många felaktigt byggda satser gör det ibland obegripligt vad skrivaren vill säga.[173] (Klju1-4-C, ext. schwed. Bewert. 1)

Während die Fehlgriffe im Text gemäß der/dem ersten Bewertenden im Text augenfällig sind, werden sie von der Lehrkraft nicht so sehr betont. Hier wird deutlich, dass die Bewertenden auch die Auswirkung der sprachlichen Mängel auf die Verständlichkeit unterschiedlich beurteilen. Die praktizierende Lehrkraft betont die guten Formulierungen im Text und kommentiert dabei nicht, dass Fehlgriffe im Bereich formaler Strukturen die Verständlichkeit beeinträchtigen, während die/der externe Bewertende im Bewerterurteil vor allem die Korrektheit und die Auswirkung syntaktischer Fehlgriffe auf die Verständlichkeit beschreibt. Ferner wird von sämtlichen Bewertenden die Erfüllung der *inhaltlichen Anforderungen* erläutert (vgl. Beispiele 7.16–7.18):

(7.16) Eleven har med alla 3 „Punkten".[174] (Klju1-4-C, Lehrkraft)

(7.17) Bra och kreativt innehåll.[175] (Klju1-4-C, ext. schwed. Bewert. 1)

(7.18) Uppgifter utförda & ett i vissa fall välutvecklat innehåll.[176] (Klju1-4-C, ext. schwed. Bewert. 1)

Die Kommentare zu inhaltlichen Aspekten in den Beispielen sind positiv wertend, auch wenn die Bewertenden unterschiedliche Perspektiven einnehmen. Sie kommentieren hierbei z. B., inwiefern alle nachgefragten Informationen in der Aufgabe behandelt werden oder inwieweit der Text einen kreativen oder entwickelten Inhalt aufweist. Zwei der Bewertenden erwähnen zudem die knappe Textlänge für die zweite Teilaufgabe. Darüber hinaus werden andere Aspekte bei der Bewertung berücksichtigt. Die praktizierende Lehrkraft und die/der zweite externe Bewertende kommentieren in positiven Worten den *Textfluss*. Einzelne Bewertende geben zudem Kommentare zur *Kohärenz* im Text, zur Umsetzung der *Textsorte*, zur *soziokulturellen Anpassung* sowie dazu,

172 „Der Text ist dennoch sehr gut formuliert und verständlich […] Einige grammatische Fehlgriffe und Rechtschreibfehler".

173 „Generell fehlerhafte Sätze. […] Viele falsch aufgebaute Sätze machen manchmal unverständlich, was der Schreiber sagen möchte".

174 „Der Schüler/die Schülerin hat die drei geforderten Punkten behandelt".

175 „Guter und kreativer Inhalt".

176 „Aufgaben gelöst und ein in gewissem Ausmaß gut entwickelter Inhalt".

inwieweit der Text *kommunikativ* ist oder nicht. Diese vereinzelten Kommentare der schwedischen Bewertenden beschreiben die Leistung in sowohl positiven als auch negativen Worten.

Zusammenfassend werden bei der Textproduktion *Klju1–4* relativ häufig die gleichen Aspekte berücksichtigt, aber vereinzelte Kommentare zu unterschiedlichen Aspekten kommen auch vor. Auch wenn die Bewertenden häufig die gleichen oder ähnlichen Dimensionen kommentieren, gewichten und interpretieren sie aber diese nicht immer auf die gleiche Weise. Dies gilt vorwiegend für Aspekte der linguistischen Kompetenz. In den Kommentaren zur inhaltlichen Erfüllung der Aufgabe haben sie teilweise unterschiedliche Perspektiven, sind sich aber dennoch ziemlich einig. Zu einem gewissen Grad kommentieren die jeweiligen Bewertenden auch unterschiedliche Aspekte, dabei indizierend, dass sie nicht immer die gleichen Gründe für die Benotung haben. Es handelt sich hierbei vorwiegend um Aspekte zur soziokulturellen Angemessenheit oder zur Umsetzung der Textsorte.

7.5 Fazit

Bezüglich der zweiten Forschungsfrage, inwieweit die schwedischen Bewertenden bei einer Beurteilung schriftlicher Kompetenz untereinander übereinstimmen, ergeben sich sowohl Gemeinsamkeiten als auch Unterschiede. Die Studie zeigt, dass die Gruppe der Deutschlehrkräfte im Vergleich zu den beiden externen Bewertenden eine Tendenz hat, in ihren Bewertungen insgesamt höhere Noten zu geben (vgl. Tab. 29). Des Weiteren deuten die Berechnungen auf eine höhere Übereinstimmung zwischen den externen schwedischen Bewertenden im Vergleich zu der Gruppe der Lehrkräfte hin (vgl. Tab. 30). In Bezug auf die Ergebnisse in Tab. 30 stellt sich somit die Frage nach der Bewerterübereinstimmung der schwedischen Bewertenden. Bei einem Vergleich fallen darüber hinaus die Konsenswerte niedriger aus als die Konsistenzwerte. Da die Berechnungen der Übereinstimmungsrate nicht im gleichen Ausmaß wie die Konsistenzwerte zufriedenstellend sind, kann angenommen werden, dass die schwedischen Bewertenden vergleichsweise in höherem Grad übereinstimmende Rangreihen von Bewertungen als exakte Übereinstimmungen erzeugen. Festzustellen ist hierbei, dass die Einstufungen der beiden externen Bewertenden in höherem Grad im Hinblick auf den Konsens und Konsistenz miteinander übereinstimmen als Vergleiche, die Einstufungen von der Gruppe der Lehrkräfte beinhalten.

Wenn sich Konsens- und Konsistenzwerte deutlich unterscheiden, deutet dies darauf, dass die untersuchten Bewertergruppen Differenzen in Bezug auf

die Beurteilerstrenge aufweisen, aber eine ähnliche Rangordnung der Textproduktionen vornehmen. Die Kreuztabellen (vgl. Tab. 31–Tab. 33) und die durchgeführte Multifacetten-Rasch-Analyse (vgl. Abb. 10) bestätigen dieses Bild: Aus den Kreuztabellen lässt sich zudem ableiten, dass die Bewertungen der externen schwedischen Bewertenden im Vergleich zu der Gruppe der Lehrkräfte generell eher niedrigere Einstufungen auf der Notenskala enthalten haben.

Die Einstufungen zum selben Text unterschieden sich demnach voneinander. Am häufigsten unterscheiden sich die Bewertungen von Textproduktionen mittlerer oder höherer Benotung. Diese Tendenz kann am deutlichsten für die Note C (mit insgesamt 6 Übereinstimmungen und 33 Nichtübereinstimmungen) beobachtet werden. Für die Note E ergeben sich gleich viele Übereinstimmungen als auch Nichtübereinstimmungen bei der Vergabe der Noten, wobei die Nichtübereinstimmungen in der Regel in den Vergleichen zwischen der Gruppe der schwedischen Deutschlehrkräfte und den beiden externen Bewertenden zu finden sind. Bei der nicht ausreichenden Note F (mit insgesamt 45 Übereinstimmungen und 9 Nichtübereinstimmungen) stimmen wesentlich häufiger die Einstufungen der schwedischen Bewertenden miteinander überein.

Aus der Multifacetten-Rasch-Analyse ergibt sich, dass die Gruppe der Lehrkräfte mit einem negativen Logitwert eine Neigung zur Milde aufweist, während die beiden externen Bewertenden mit positiven Logitwerten leichte Strenge-Tendenzen vorweisen. Die Gruppe der Lehrkräfte zeigt somit im Vergleich zu den beiden externen schwedischen Bewertenden eine Tendenz, die Textproduktionen ihrer eigenen Lernenden generell milder zu bewerten. Insgesamt ist auch durch die Multifacetten-Rasch-Analyse festzustellen, dass die Differenzen hinsichtlich der Bewerterstrenge zwischen den beiden externen Bewertenden geringer sind als im Vergleichen zu der Gruppe der schwedischen Lehrkräfte. Allerdings ist sowohl die Anzahl der benoteten Texte als auch der Bewertenden im Datensatz relativ begrenzt, weshalb die Ergebnisse mit Vorsicht interpretiert werden sollten.

Zusammenfassend enthalten die Bewerterurteile der schwedischen Bewertenden sowohl Gemeinsamkeiten als auch Unterschiede im Hinblick auf die in den Textproduktionen beachteten Aspekte und darauf, wie diese eingeschätzt werden. Auch wenn die Bewertenden bei der Benotung nicht übereinstimmen, scheinen sie relativ häufig ähnliche Aspekte zu bewerten, mitunter gewichten sie jedoch diese Aspekte unterschiedlich, was als möglicher Grund für die divergierende Benotung erscheint. Zu bemerken hierbei ist, dass Merkmale in den Schülerleistungen, die zu denselben Bewertungsdimensionen gehören, bei der Bewertung zu unterschiedlicher Gewichtung führen können. Aus dem qualitativen Vergleich zwischen den Bewerterurteilen mit divergierenden Benotungen

kann zudem geschlossen werden, dass von den jeweiligen schwedischen Bewertenden gelegentlich auch unterschiedliche Aspekte in den Textproduktionen berücksichtigt werden. Beispiele dieser Unterschiede sind die Bewertungen von Aspekten zu *inhaltlicher Aufgabenerfüllung* und *Aspekten der linguistischen Kompetenz*. Auch bei Textproduktionen, die dieselbe Benotung erhalten haben, beachten die Bewertenden manchmal verschiedene Aspekte oder gewichten dieselben Aspekte unterschiedlich, was die Frage aufkommen lässt, ob sie gelegentlich aus verschiedenen Gründen zu denselben Noten kommen. Es scheint somit bisweilen unterschiedliche Gründe für sowohl divergierende als auch übereinstimmende Benotungen unter den schwedischen Bewertenden zu geben.

Wenn die Bewertenden bei der Benotung übereinstimmig sind, entsteht ein etwas anderes Bild. Wenn sämtliche schwedischen Bewertenden einer Schülerleistung eine nicht ausreichende Note F geben, berücksichtigen sie in höherem Ausmaß die gleichen Aspekte. Generell könnte dementsprechend angenommen werden, dass die Bewertenden häufiger ähnliche Gewichtungen und Interpretationen für die Mindestanforderungen in den Schülerleistungen haben. Es handelt sich dabei um ähnliche Interpretationen zur Bedeutung von *Textlänge* (in der Aufgabe angegeben), der Beherrschung grundlegender Kenntnisse in den Bereichen *Grammatik* und *Wortschatz* und deren Auswirkung auf die *Verständlichkeit* sowie der Beeinträchtigung der Note, wenn eine Aufgabe nicht gelöst ist. Im mittleren Bereich scheinen die schwedischen Bewertenden dahingegen bei der Benotung von Textproduktionen im Hinblick darauf, *welche* Aspekte sie beurteilen und *wie* sie diese gewichten, etwas weniger miteinander übereinzustimmen. Es handelt sich u. a. darum, in welchem Ausmaß Fehlgriffe das Verständnis beeinträchtigt, welche Merkmale oder Kriterien bei den inhaltlichen Anforderungen erfüllt werden müssen, und inwiefern Aspekte zur soziokulturellen Angemessenheit oder zur Umsetzung der Textsorte überhaupt kommentiert werden sollten.

Abschließend zeigen die quantitativen und die qualitativen Analysen auf eine mangelnde Bewerterübereinstimmung bei der Bewertung durch die schwedischen Bewertenden. Zu beachten ist, dass die Lehrkräfte eine starke Tendenz zur Milde aufweisen im Vergleich zu den beiden externen Bewertenden, die dahingegen eine leichte Tendenz zur Strenge zeigen. Des Weiteren geht aus den Analysen hervor, dass die Bewertenden bei der Bewertung sowohl ähnliche als auch unterschiedliche Aspekte beachten und dass sie gelegentlich die gleichen Aspekte unterschiedlich gewichten. Dies scheint zudem sowohl zu unterschiedlicher Benotung als auch zur Vergabe derselben Note führen zu können.

8. Analyse der Beziehung zum B1-Niveau

Dieses Kapitel befasst sich mit den Ergebnissen der GER-Bewertungen hinsichtlich eines B1-Niveaus und setzt die Bewertungen nach schwedischen Bildungsstandards mit den GER-Bewertungen in Verbindung. Hierbei wird die dritte Fragestellung der vorliegenden Arbeit untersucht: *In welcher Beziehung stehen Bewertungen von Textproduktionen schwedischer Schülerinnen und Schüler auf den Fremdsprachenstufen* Tyska 3, Tyska 4 *und* Tyska 5 *des schwedischen Bildungssystems zu Bewertungen der schriftlichen Sprachkompetenz auf einem erfüllten B1-Niveau des GER?* Die Referenzniveaus des GER funktionieren zunehmend als Bezugspunkt für fremdsprachliche Kompetenz und dienen auch für das schwedische Stufenmodell als Referenzpunkt für erreichte Sprachkompetenzen am Ende der jeweiligen Fremdsprachenstufen. Ausgangspunkt für ein erreichtes Niveau B1.2 des GER sind im schwedischen System die Mindestanforderungen für eine ausreichende E-Note auf *Tyska 5* (vgl. Kap. 2.4.2). Um dies zu untersuchen, wurden Lernproduktionen des schriftlichen Ausdrucks im Hinblick auf ein erfülltes B1-Niveau in einem ersten Schritt von schwedischen Bewertenden beurteilt und in einem zweiten Schritt von jeweils zwei unabhängigen GER-Bewertenden nachbewertet. Die GER-Bewertenden haben die jeweiligen Textproduktionen ohne jegliche Kenntnisse über die ursprüngliche Evaluation und die jeweiligen Fremdsprachenstufen der Probanden im Hinblick auf das B1-Niveau bewertet.

In diesem Kapitel geht es im Folgenden zunächst um die Ergebnisse der schwedischen Bewertungen und darum, wie diese zu einem erreichten bzw. nicht-erreichten B1-Niveau in Verhältnis stehen. Ein Ziel der Untersuchung ist es dementsprechend, die Testergebnisse der schwedischen Bewertungen gegen das von Skolverket auf *Tyska 5* angestrebte GER-Niveau B1.2 zu überprüfen. Zunächst wird ausgewertet, wie Textproduktionen, die die Anforderungen eines B1-Niveaus bei dem schriftlichen Test erfüllt haben, sich auf die drei untersuchten Fremdsprachenstufen verteilen (Kap. 8.1). Anschließend wird untersucht, inwiefern bei *Tyska 5* Leistungen mit mindestens einer ausreichenden Note E dem von der schwedischen Schulbehörde intendierten GER-Niveau BI entsprechen (Kap. 8.2). Des Weiteren wird ermittelt, inwiefern Schülerleistungen, die die Anforderungen eines B1-Niveaus bei dem schriftlichen Test erfüllt haben, zu einer bestimmten Benotung oder einem Fremdsprachenniveau des schwedischen Systems in Verbindung gesetzt werden können. Darüber hinaus werden Korrelationen zwischen den jeweiligen Ergebnissen berechnet, um ein

vielfältigeres Bild der Beziehung zwischen der Bewertung fremdsprachlicher Lernproduktionen nach schwedischen Bildungsstandards und einem erreichten GER-Niveau BI zu erhalten (Kap. 8.3). Danach wird ein Vergleich der Bewertungen zweier grenzwertiger Lernproduktionen unternommen. Gemäß den GER-Bewertenden erreichen diese beiden Leistungen die Anforderungen eines B1-Niveaus, während zwei der schwedischen Bewertenden den Texten eine unbefriedigende Note geben (Kap. 8.4). Abschließend werden die Befunde dieser Analysen kurz zusammengefasst (Kap. 8.5).

8.1 Deskriptive Statistik hinsichtlich des Niveaus B1

Zur Auswertung der Beziehung zwischen Testergebnissen schwedischer Bewertungen und GER-Bewertungen schriftlicher Kompetenz wurde der vom Goethe-Institut entworfene, im Rahmen des *Goethe-Zertifikats B1* eingesetzte, Prüfungsteil *Schreiben* verwendet. In der Prüfung können maximal 100 Punkte erreicht werden. Um bei der Prüfung *Zertifikat B1* ein B1-Niveau der schriftlichen Kompetenzen zu erreichen, müssen die Textproduktionen eine Punktzahl von mindestens 60 Punkten erreicht haben. Die beiden GER-Bewertenden stimmen beim Feststellen eines erreichten B1-Niveaus für die Lernproduktionen bis auf eine Ausnahme überein, was die Reliabilität der Ergebnisse bei der GER-Bewertung stärkt. Es handelt sich dabei um eine Textproduktion aus *Tyska 3*, die in der Beurteilung des ersten GER-Bewertenden eine Punktzahl von 59,5 erreicht hat (ein beinahe erreichtes B1-Niveau), während die/der zweite GER-Bewertende für diese Textproduktion eine Punktzahl von 70,5 ermittelt hat. Zur Ermittlung eines Gesamtergebnisses wird ein arithmetisches Mittel berechnet. Das Gesamtergebnis für die GER-Bewertung zeigt daher, dass jene Textproduktion mit einer durchschnittlichen Punktzahl von 65/100 das B1-Niveau erreicht hat.

Die deskriptive Statistik für die GER-Bewertungen, basierend auf den Gesamtergebnissen für den Prüfungsteil *Schreiben*, ist getrennt nach Fremdsprachenstufen aus Tab. 35 zu erlesen:

Tab. 35: *Deskriptive Statistik (Extremwerte, Mittelwerte, Mediane und Standardabweichungen) für die GER-Bewertungen nach Fremdsprachenstufe*

Sprachstufe	*N*	*Minimum*	*Maximum*	*Mittelwert*	*Median*	*Std*
Tyska 3	20	10	97	50,50	51,00	27,78
Tyska 4	20	11	98	59,15	69,50	31,31
Tyska 5	20	13	100	78,80	86,00	23,25
Gesamt	**60**	**10**	**100**	**62,82**	**68,83**	**27,45**

Tab. 35 stellt die Gesamtergebnisse für die untersuchten Lernproduktionen dar. Die niedrigste gegebene Punktzahl liegt bei 10 (*Tyska 3*), die höchste bei 100 Punkten (*Tyska 5*). Die Minimal- bzw. Maximalwerte der jeweiligen Stufen zeigen jedoch, dass innerhalb sämtlicher Stufen leistungsschwächere bzw. leistungsstärkere Textproduktionen zu finden sind. Bei der Betrachtung der Mittelwerte wird deutlich, dass der Mittelwert für die höchste Stufe, *Tyska 5*, über der Bestehensgrenze liegt, während diese Werte für die beiden anderen Stufen, *Tyska 3* und *Tyska 4*, unter der geforderten Grenze von 60 Punkten liegen. Allerdings zeigt der Medianwert von 69,50, dass über die Hälfte der Textproduktionen auf *Tyska 4* das B1-Niveau erreichen. Darüber hinaus ist der Medianwert auf *Tyska 5* mit 86 Punkten sehr hoch. Der Median besitzt die Eigenschaft, gegen Ausreißer robuster zu sein und kann daher in dieser Studie relevant sein. Die Standardabweichungen liegen zwischen den Werten 23,25 und 31,31, wobei *Tyska 3* und *Tyska 4* die höheren Werte zu verzeichnen haben. Die beiden Stufen zeigen dementsprechend eine höhere Streuung der Punktzahlen innerhalb der Stufe, während die Ergebnisse der GER-Bewertungen auf *Tyska 5* im Vergleich zu den anderen beiden Stufen nicht so weit voneinander abweichen.

In Abb. 11 wird zunächst dargestellt, wie sich die untersuchten Schülerleistungen im Bereich schriftlicher Sprachfertigkeit bezüglich der erreichten Punktzahlen verteilen. Hierbei werden die Gesamtergebnisse der GER-Bewertungen (0–100 Punkte) getrennt nach Fremdsprachenstufe in Form eines Boxplot-Diagramms abgebildet:

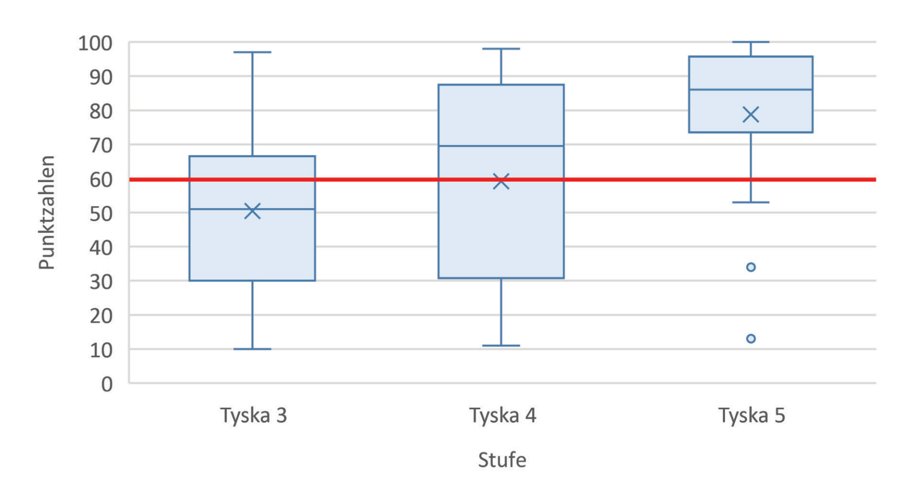


Abb. 11: Boxplot-Diagramm: Verteilung der Lernproduktionen auf die Fremdsprachenstufen nach Punktzahlen bei der GER-Bewertung (*N* = 60)

In der Abbildung zeigt die horizontale Achse die jeweiligen Fremdsprachenstufen und die vertikale Achse bezieht sich auf die jeweiligen Punktzahlen der Probanden (Wertebereich 1–100 Punkte). Die durchgehende rote Linie in der Abbildung markiert die Bestehensgrenze bei 60 Punkten zwischen einem erreichten bzw. nicht-erreichten B1-Niveau. Ein Boxplot-Diagramm ist von den zugrundeliegenden Daten abhängig. Der Median ist im Diagramm mit einer durchgehenden Linie visualisiert und der Mittelwert mit einem Kreuz veranschaulicht. Die Boxen der jeweiligen Fremdsprachenstufen zeigen, in welcher Spanne sich die mittlere Hälfte aller Textproduktionen befindet. Aus der Abbildung wird ersichtlich, dass die Textproduktionen auf *Tyska 3* und *Tyska 4* sich durch eine größere Streuung auszeichnen als die auf *Tyska 5*, die weitgehend sehr hoch liegen. Es wird hierbei auch deutlich, dass die große Mehrheit der Textproduktionen, die zur höchsten Stufe *Tyska 5* gehören, deutlich über der Bestehensgrenze liegt. Im Boxplot sind dennoch auf *Tyska 5* auch zwei Ausreißer zu erkennen, die mit den niedrigen Punktzahlen 34 bzw. 13 von den anderen Textproduktionen abweichen. Umgekehrt befindet sich die große Mehrheit der Textproduktionen aus der Stufe *Tyska 3* unter dieser Grenze, auch wenn einzelne Textproduktionen hohe Punktzahlen aufzeigen können. Die Textproduktionen auf *Tyska 4* sind sowohl über als unter der Bestehensgrenze zu finden. Der Median liegt jedoch, wie bereits in Tab. 35 ersichtlich wurde, jedoch bei *Tyska 4* deutlich über der Bestehensgrenze.

Die Verteilung der einzelnen Lernproduktionen nach einem erreichten bzw. nicht-erreichten B1-Niveau für die Fremdsprachenstufen *Tyska 3*, *Tyska 4* und *Tyska 5* lässt sich in Tab. 36 ablesen:

Tab. 36: *Verteilung der GER-Bewertungen hinsichtlich des Sprachniveaus B1 (Anzahl und Prozent angegeben)*

Fremdsprachenstufe	*N*	*nicht-erreichtes B1-Niveau*		*erreichtes B1-Niveau*	
		N	*%*	*N*	*%*
Tyska 3	20	13	68	7	32
Tyska 4	20	9	45	11	55
Tyska 5	20	4	20	16	80
Gesamt	**60**	**26**	**43**	**34**	**56**

Tab. 36 zeigt die Anzahl und wie viel Prozent der Schülerleistungen auf den jeweiligen Fremdsprachenstufen das B1-Niveau bei dem schriftlichen Test erreicht haben.[177] Aus der Tabelle ist zu erkennen, dass mehr als die Hälfte der

177 Es ist bei den Prozentberechnungen zu beachten, dass die Daten nach einer bewussten Textauswahl entstanden sind, nicht ausgehend von der Gesamtpopulation.

Textproduktionen auf einem B1-Niveau eingestuft wurden (34/60). Dies betrifft nicht überraschend vor allem die Lernproduktionen aus *Tyska 5*. Auf dieser Stufe haben 16/20 der Texte bei der Bewertung ein erfülltes B1-Niveau erreicht. Die große Mehrheit der Schülerleistungen auf *Tyska 5* hat dementsprechend das von der schwedischen Schulbehörde angesprochene Niveau erreicht. Darüber hinaus haben auch 11/20 der getesteten Textproduktionen auf *Tyska 4* schriftliche Kompetenzen auf einem B1.2-Niveau gezeigt und somit haben in etwa die Hälfte der getesteten Probanden auf *Tyska 4* das Mindestniveau im Bereich Schreiben in Bezug auf das GER-Niveau für die Stufe übertroffen. Zu bemerken ist zudem, dass 7/20 der untersuchten Leistungen im Kurs *Tyska 3* die schriftliche Prüfung auf dem B1.2-Niveau bewältigt haben. In einem dieser Fälle hat jedoch der eine GER-Bewertende, wie bereits oben erwähnt, eine Textproduktion als nicht bestanden bewertet, d. h. unter der Bestehensgrenze von 60 Punkten.

Insgesamt haben 26/60 der getesteten Schülerproduktionen das B1-Niveau nicht erreicht. Von denjenigen, die das B1-Niveau nicht erreicht haben, befinden sich die meisten auf der im Datensatz niedrigsten Stufe *Tyska 3*: hier sind 13/20 der untersuchten Schülerleistungen unter dem Niveau B1 eingestuft worden. Nicht erreicht haben es auch 9/20 auf *Tyska 4* und 4/20 auf *Tyska 5*. Letztere haben somit das erwartete Niveau für diese Stufe, wenn man sich nach dem GER-Niveau B1 orientiert, nicht erreicht.

Des Weiteren ist auch von Relevanz, welche Ergebnisse die untersuchten Leistungen auf den jeweiligen Fremdsprachenstufen erhalten haben. Tab. 36 zeigte die Ergebnisse der GER-Bewertungen nur nach den Kategorien *erreichtes B1-Niveau* bzw. *nicht-erreichtes B1-Niveau*. Nach den Anweisungen für den schriftlichen Test werden die Textproduktionen jedoch mit Punktzahlen und Prädikaten dokumentiert (vgl. Kap. 5.2). Wie gut die untersuchten Textproduktionen bei der schriftlichen Prüfung abgeschnitten haben, ist in Tab. 37 nach diesen Punktzahlen und Prädikaten zusammengefasst:

Tab. 37: *Verteilung der Bewertungen nach den von den GER-Bewertenden ermittelten Punktzahlen auf die jeweiligen Fremdsprachenstufen (N = 60)*

Fremdsprachenstufe	*0–60 nicht erreicht*	*60–69 ausreichend*	*70–79 befriedigend*	*80–89 gut*	*90–100 sehr gut*
Tyska 3	13	3	–	1	3
Tyska 4	9	1	2	4	4
Tyska 5	4	–	3	5	8

Die Ergebnisse zeigen getrennt nach Fremdsprachenstufen insgesamt ein gemischtes Bild. Aus Tab. 37 ist zu entnehmen, dass Schülerleistungen mit hohem Kompetenzniveau gemäß den beiden GER-Bewertenden bereits auf *Tyska 3* nachzuweisen sind: Auf *Tyska 3* haben vier der zwanzig Schülerleistungen Punktzahlen zwischen 80 und 100 erhalten. Die große Mehrheit der Leistungen auf *Tyska 3* hat dahingegen entweder das Niveau B1 sehr knapp oder gar nicht erreicht. Auf *Tyska 3* befinden sich aber kaum Schülerleistungen im mittleren Bereich – sie scheinen entweder die Aufgabe *gut* bis *sehr gut* bewältigt zu haben oder knapp bestanden. Diese Ergebnisse deuten somit darauf hin, dass die Aufgabe für viele der Schülerinnen und Schüler jener Stufe auf einem zu hohen Niveau war. Auf *Tyska 4* ist eine größere Anzahl von Textproduktionen, die ein gutes bzw. sehr gutes Niveau nachgewiesen haben, zu finden (8/20).

Des Weiteren geht aus Tab. 37 hervor, dass die Anzahl von Lernproduktionen mit höheren Punktzahlen mit jeder Fremdsprachenstufe steigt. Der Trend ist dementsprechend deutlich: die größte Anzahl der Schülerleistungen, die den schriftlichen Test *gut* oder *sehr gut* bewältigt haben, befindet sich auf *Tyska 5*. Ganze 13 von 20 Schülerleistungen auf *Tyska 5* erhalten Punktzahlen, die als *gut* oder *sehr gut* zu interpretieren sind. Die große Mehrheit der Texte auf dieser Stufe zeigt hier auf ein sehr hohes Niveau hinsichtlich der schriftlichen Sprachkompetenz der Lernenden.

8.2 Auswertung der Orientierung am GER

Das siebenstufige Modell des schwedischen Systems nimmt die GER-Niveaus als Ausgangspunkt. Hierbei beziehen sich die Mindeststandards der jeweiligen Fremdsprachenstufen auf ein erreichtes Niveau oder gegebenenfalls Subniveau des GER (vgl. Kap. 2.4). Im vorliegenden Abschnitt wird die empirische Zuordnung von Textproduktionen auf der Fremdsprachenstufen *Tyska 5* des schwedischen Bildungssystems zur GER-Stufe B1 untersucht. Die Einstufungen, die die jeweiligen schwedischen Bewertenden, d. h. die Gruppe der Lehrkräfte und die beiden externen Bewertenden, erteilt haben und die Gesamtergebnisse der GER-Bewertung sind einander in Tab. 38 gegenübergestellt:

Tab. 38: *Ergebnisse der Bewertungen und Niveauzuordnung für die Textproduktionen auf Tyska 5 (N = 20)*

Schülerleistung	*Gruppe der Lehrkräfte*	*ext. schwed. Bewert. 1*	*ext. schwed. Bewert. 2*	*GER-Bewertung (Gesamtergebnis)*	*geschätztes GER-Niveau*
Kasv3–5	A	A	C	100	B1
Ihsu1–5	A	C	A	100	B1
Tect4–5	A	B	C	100	B1
Hjjg6-5	A	C	B	99	B1
Ekls1–5	A	D	C	91	B1
Smvl2–5	C	B	B	97	B1
Hchg2–5	C	E	D	92	B1
Kinv5-5	C	D	D	91	B1
Eles2–5	C	E	E	86	B1
Kpnu28-5	C	F	E	75	B1
Hvbg3–5	D	E	E	85	B1
Hjlg4–5	D	E	E	81	B1
Twpt3–5	D	E	E	75	B1
Ilgs11-5	E	D	D	86	B1
Soal1-5	E	F	F	13	A2
Slsk1–5	F	F	E	87	B1
Pnmj1–5	F	F	E	73	B1
Rdsv1–5	F	F	F	58	A2
Rjrv2–5	F	F	F	53	A2
Sjel3–5	F	F	F	34	A2

Für die Fremdsprachenstufe *Tyska 5* gilt als angestrebtes Niveau das GER-Niveau B1.2, was wiederum im schwedischen System bedeutet, dass bei *Tyska 5* die Voraussetzungen für eine Leistung mit mindestens ausreichender E-Note sich am GER-Niveau B1 orientieren (vgl. Skolverket 2011b). In Tab. 38 sind die Schülerleistungen nach der initialen Noteneinstufung F–A durch die Gruppe der schwedischen Lehrkräfte angeordnet und deren nach den GER-Bewertungen geschätzte GER-Stufen aufgeführt. Die zweite Spalte der Tabelle zeigt die Ergebnisse dieser Bewertungen durch die Gruppe der Lehrkräfte, gefolgt von den Bewertungen der beiden externen schwedischen Bewertenden in der dritten und vierten Spalte. Aus Spalte fünf wird das Gesamtergebnis der beiden GER-Bewertungen ersichtlich. Die letzte Spalte bildet eine Interpretation des geschätzten GER-Niveaus der jeweiligen Schülerleistungen nach der GER-Bewertung. Die grün unterlegten Felder in Tab. 38 erweisen, welche

Lernproduktionen mindestens eine ausreichende Note E nach den schwedischen Bildungsstandards oder die Bestehensgrenze für ein B1-Niveau im Prüfungsteil des schriftlichen Ausdrucks erreicht haben (eine Punktzahl von mindestens 60 Punkten). Die Texte, für die ein erreichtes B1-Niveau nicht vergeben wurde, sind auf GER-Niveau A2 eingestuft worden, da angenommen werden kann, dass die Lernenden dieses Niveau erreicht haben.

Wie sich Tab. 38 entnehmen lässt, erhalten sämtliche Textproduktionen mit der Note A (oder B) im Datensatz in den GER-Bewertungen das B1-Niveau. Diese haben alle Punktzahlen über 90 Punkte, wobei ganze drei Texte die maximale Anzahl an Punkten (100) erhalten haben. Diese hohen Punktzahlen könnten darauf hindeuten, dass diese Lernproduktionen auch auf höhere GER-Niveaus eingestuft werden könnten. Zu bemerken ist aber, dass einige dieser Leistungen mit Punktzahlen über 90 Punkte von den schwedischen Bewertenden auch Noten zwischen E und C erhalten haben.

Überdies erhalten die Textproduktionen mittlerer Benotung (mit mindestens eine Note E) mit Ausnahme der Schülerleistung *Soal1–5* eine Punktzahl von 73 Punkten und erreichen damit auch die Bestehensgrenze für eine GER-Stufe B1 erreicht. Diese Leistungen scheinen nach der GER-Bewertung auf einem relativ stabilen Niveau B1 zu liegen. Es handelt sich bei der Ausnahme um eine Schülerleistung (*Soal1–5*), die von der eigenen Lehrkraft eine ausreichende E-Note erhalten hat, wobei die beiden externen Bewertenden diese Leistung eine nicht ausreichende Note F erteilt haben. Diese erwähnte Textproduktion wurde wie bei der externen Bewertung auch von den GER-Bewertenden als ein nicht erreichtes B1-Niveau betrachtet und hat bei der GER-Bewertung nur 13 Punkte erhalten.

In Übereinstimmung mit der Intention der schwedischen Schulbehörde erreichen häufig Textproduktionen auf *Tyska 5* ohne eine ausreichende Note Enicht das B1-Niveau. Von Interesse sind aber die Fälle, bei denen die Bewertenden im Hinblick auf eine unbefriedigende Note F bzw. ein erreichtes B1-Niveau uneinig waren. Dies gilt zum einen für die Schülerleistung *Kpnu28-5*, die von der/dem ersten externen Bewertenden eine unbefriedigende Note F erhalten hat, aber sowohl in den übrigen schwedischen Bewertungen als auch in den GER-Bewertungen als ausreichend für das Niveau beurteilt wurde. Bei zwei anderen Schülerleistungen, *Slsk1–5* und *Pnmj1–5*, handelt es sich um jeweils zwei unbefriedigende Noten, sowohl aus der Gruppe der Lehrkräfte als auch von der/dem ersten schwedischen Bewertenden. Vom zweiten schwedischen Bewertenden haben diese Schülerleistungen allerdings eine ausreichende E-Note bekommen und sie haben nach der GER-Bewertung die Anforderungen für das B1-Niveau erfüllt. Diese Schülerleistungen liegen womöglich an der

Grenze gemäß den schwedischen Bewertenden, scheinen aber nach den GER-Bewertungen auf einem ziemlich stabilen B1-Niveau zu liegen.

Auch wenn die Bewertungen im niedrigeren Bereich ein manchmal gemischtes Bild gezeichnet haben, wird aus Tab. 38 zudem deutlich, dass die Übereinstimmungen hinsichtlich der Benotung am unteren Ende zwischen den schwedischen Bewertenden etwas größer sind. Deutlich wird somit auch die Streuung der Notengebung der schwedischen Bewertenden im Vergleich mit den Ergebnissen der GER-Bewertenden, vor allem für die mittleren und in gewissem Ausmaß auch für die höheren Noten. Das Verhältnis zwischen den Bewertungen der beiden Bewertergruppen wird in Abschnitt 8.3 weiter behandelt.

8.3 Zum Verhältnis schwedischer Bewertungen und GER-Bewertungen

Der folgende Abschnitt befasst sich zunächst mit dem Verhältnis der schwedischen Bewertungen und der GER-Bewertungen zueinander. Hierbei wird auf die Frage nach der Verbindung zwischen den einzelnen Notenstufen im schwedischen System und einem erreichten B1-Niveau fokussiert. Um das Zusammenspiel zwischen einer Benotung fremdsprachlicher Lernproduktionen nach schwedischen Bildungsstandards und der GER-Bewertung untersuchen zu können, werden die Ergebnisse der beiden Bewertungsverfahren korreliert. Es handelt sich hierbei sowohl um Korrelationen zwischen den jeweiligen Gesamtergebnissen als auch um Vergleiche mit den Aspektbewertungen bei der GER-Bewertung.

Erforderliche Kompetenzen für eine zweite Fremdsprache auf der dritten Stufe des schwedischen Systems, *Tyska 3*, orientieren sich in Richtung eines erfüllten A2-Niveaus des GER. Dies bedeutet wiederum, dass ein erreichtes A2-Niveau als Referenzpunkt für das Mindestniveau einer ausreichenden Note E auf *Tyska 3* angenommen wird. Bei einer Bewertung sollte aber nicht nur die Orientierung am Mindestniveau definiert werden; es kann dabei auch wichtig sein, einen Referenzpunkt zum Niveau für die höheren Noten zu definieren (vgl. North 2014: 208). Wenn folglich ein erreichtes A2-Niveau für die niedrigste Note E verlangt wird, kann dann angenommen werden, dass Schülerleistungen, die die höchste Note A erhalten, ein B1-Niveau im Hinblick auf die schriftliche Kompetenz erreicht haben? Für die tentative Zuordnung zum Referenzniveau B1 wurden die Ergebnisse der Lernproduktionen aus den drei Fremdsprachenstufen untersucht, um Tendenzen im Material nachgehen zu können. Die ursprüngliche Benotung der Lernproduktionen durch die Gruppe

der schwedischen Lehrkräfte, die der Auswahl von Texten zugrunde liegt, bildet den Ausgangspunkt für diesen Vergleich. Tab. 39 zeigt die Verteilung der Bewerterurteile auf die jeweiligen Fremdsprachenstufen, die, basierend auf der GER-Einstufung, das B1-Niveau erreicht bzw. nicht erreicht haben:

Tab. 39: *Verteilung der Textproduktionen eines erreichten B1-Niveaus auf die jeweiligen Fremdsprachenstufen nach der Benotung der schwedischen Lehrkräfte*

Fremdsprachenstufen	*F*	*E (D)*	*C*	*(B) A*
Tyska 3				
Tyska 4				
Tyska 5				

Die grüne Unterlegung der Felder in der Tabelle oben zeigt an, inwiefern eine Mehrheit der Lernproduktionen auf den jeweiligen Notenstufen die Bestehensgrenze für ein B1-Niveau im Prüfungsteil des schriftlichen Ausdrucks erreicht hat. Tab. 39 ist somit zu entnehmen, dass Schülerleistungen mit hohem Kompetenzniveau gemäß den GER-Bewertungen bereits auf *Tyska 3* und *Tyska 4* nachzuweisen sind. Dies wurde schon daraus ersichtlich, dass 3/20 von den Schülerleistungen auf *Tyska 3* und 4/20 auf *Tyska 4*, alle mit hohen Noten, sogar Punktzahlen zwischen 90 und 100 erhalten (vgl. Tab. 37). Hoch benotete Textproduktionen jener Stufen scheinen somit auf einem B1.2-Niveau zu liegen, da sämtliche mit A oder B benoteten Schülerleistungen auf Tyska 3 und Tyska 4 die für das B1-Niveau bei diesem schriftlichen Test erforderlichen Kompetenzen erreicht haben. Inwiefern die Schülerleistungen auch das Mindestsprachniveau für B2 oder noch höhere Niveaustufen des GER erreichen würden, ist jedoch im Rahmen dieser Studie nicht untersucht worden.

Einen Hinweis auf höhere Kompetenzen können uns lediglich die Punktzahlen der GER-Bewertungen geben. Diese wurden auf der Skala von nicht erreicht bis sehr gut erteilt (0–100 Punkte). Die Punktzahlen der Textproduktionen mit den höchsten Noten A und B der untersuchten Stufen befinden sich weitestgehend in einer Spannbreite von 65 bis 97 Punkten auf Tyska 3 bzw. 86 bis 98 Punkten auf *Tyska 4*. Dabei ist wahrzunehmen, dass die Spannbreite auf *Tyska 3* größer ist und dass das sprachliche Niveau, um auf *Tyska 4* die höchsten Noten zu erhalten, somit höher zu liegen scheint (vgl. hierzu auch Tab. 37).

Des Weiteren haben die Textproduktionen auf *Tyska 4*, die von den schwedischen Lehrkräften mit der mittleren Benotung C benotet worden sind, ebenfalls in den GER-Bewertungen das B1-Niveau erreicht. Diese Tendenz ist dennoch nicht so eindeutig wie das Verhältnis zwischen den höchsten Noten und dem

B1-Niveau, da die C-Note von den beiden schwedischen externen Bewertenden nur in einem Fall bestätigt werden konnte. Darüber hinaus kann gezeigt werden, dass die große Mehrheit der Textproduktionen auf *Tyska 5*, wie auch aus der Tab. 36 ersichtlich, das intendierte GER-Niveau B1 erreicht. Dies betrifft Textproduktionen mit der Benotung E–A, die damit die Anforderungen für den Kurs im Hinblick auf die schriftliche Produktion erfüllt haben.

Im Folgenden wird zudem das Verhältnis sämtlicher Bewertungen der beiden Bewertergruppen unabhängig von Fremdsprachenstufen untersucht: In welcher Beziehung Testergebnisse von Bewertungen nach schwedischen Bildungsstandards und Bewertungen hinsichtlich des GER-Niveaus B1 zueinander stehen, kann durch Berechnungen von Korrelationen untersucht werden. Hierzu dienen sowohl die Ergebnisse der jeweiligen Bewertungen auf *Tyska 3*, *Tyska 4* und *Tyska 5* von den schwedischen Bewertenden als auch das Gesamtergebnis der GER-Bewertungen. Die Gesamtpunktzahlen der jeweiligen Textproduktionen aus der GER-Bewertung wurden daher mittels einer Korrelationsanalyse (Spearman's Rho) mit den Noten der einzelnen schwedischen Bewertungen verglichen, um die Beziehungen zwischen den Urteilen zu untersuchen. Die Ergebnisse dieser Korrelationsanalyse sind in Tab. 40 dargestellt:

Tab. 40: Korrelationen zwischen den Bewertungen der schwedischen Bewertenden und dem Gesamtergebnis der GER-Bewertung (Spearman's Rho)

Bewertungen	*Gruppe d. Lehrkräfte*	*ext. schwed. Bewert. 1*	*ext. schwed. Bewert. 2*
GER-Bewertung	.787	.792	.871

Wie aus der Tabelle ersichtlich, weisen die Ergebnisse auf relativ starke Korrelationen zwischen dem Gesamtergebnis der GER-Bewertung und den schwedischen Bewertungen hin. Die Korrelationskoeffizienten reichen bei Spearman's Rho von $r = .787$ bzw. .792 ($p < 0.01$) für die Gruppe der Lehrkräfte und den ersten externen schwedischen Bewertenden bis $r = .871$ ($p < 0.01$) für den zweiten schwedischen Bewertenden, was als eine starke Korrelation zu betrachten ist. Die GER-Bewertungen korrelieren damit stärker mit der Benotung des zweiten schwedischen Bewertenden als mit der Gruppe der Lehrkräfte und der/dem ersten schwedischen Bewertenden. Korrelationen lassen keine Aussagen über kausale Wirkzusammenhänge zu, sie können aber Hinweise auf ein Verhältnis zwischen zwei Variablen geben. Diese Ergebnisse deuten darauf hin, dass von den beiden Bewertergruppen ein ähnliches Konstrukt beurteilt wurde.

In welchem Verhältnis die schwedischen Bewertungen zu den verschiedenen Aspektbewertungen stehen, kann möglicherweise Tendenzen bei einer

Beurteilung enthüllen. Bei der GER-Bewertung haben die Bewertenden unterschiedliche Bewerteraspekte, die im Bewertungsraster explizit zu finden sind, auf einer fünfgradigen Skala mit Punktzahlen evaluiert. Mit diesen Aspektbewertungen der Bewertungsdimensionen *Erfüllung*, *Kohärenz*, *Wortschatz* und *Strukturen* als Grundlage wurden hier Berechnungen von Korrelationen (Spearman's Rho) durchgeführt, um die Beziehung zwischen den schwedischen Bewertungen und den jeweiligen Aspektbewertungen der GER-Bewertenden zu untersuchen. Die Punktzahlen der jeweiligen Teilaspekte in den Urteilen der beiden GER-Bewertenden wurden daher mit den Noten der jeweiligen holistischen schwedischen Bewertungen mittels der Korrelationsanalyse verglichen, siehe Tab. 41:

Tab. 41 : Korrelationen zwischen schwedischen Bewertungen und den GER-Bewertungen hinsichtlich einzelner Bewerteraspekte (Spearman's Rho)

Aspektbewertung	*Gruppe der Lehrkräfte*		*ext. schwed. Bewert. 1*		*ext. schwed. Bewert. 2*	
	GER 1	*GER 2*	*GER 1*	*GER 2*	*GER 1*	*GER 2*
Erfüllung	.719	.744	.707	.718	.783	.769
Kohärenz	.775	.765	.763	.749	.839	.795
Wortschatz	.788	.800	.787	.795	.851	.850
Strukturen	.803	.804	.793	.786	.860	.859

Wie bei den Korrelationsberechnungen zwischen den schwedischen Bewertungen und dem gesamten GER-Ergebnis deuten auch hier die Korrelationen mit den Teilaspekten insgesamt auf ein starkes Verhältnis (zwischen $r = .707$ und $r = .860$, $p < 0.01$). Insbesondere die Bewerteraspekte *Wortschatz* und *Strukturen* korrelieren auf einem starken Niveau mit den Bewertungen der schwedischen Bewertenden (für *Wortschatz* zwischen $r = .788$ und .851, $p < 0.01$ für *Strukturen* zwischen $r = .786$ und .860, $p < 0.01$). Der Bewerteraspekt *Kohärenz* korreliert ebenfalls signifikant mit den schwedischen Bewertungen (zwischen $r = .749–.839$, $p < 0.01$). Etwas auffallend ist aber, dass das Kriterium zur *Erfüllung* im Vergleich zu den anderen Teilaspekten etwas schwächer mit den Bewertungen sämtlicher schwedischer Bewertenden korreliert (zwischen $r = .707–.783$, $p < 0.01$). Die Korrelationswerte sind bei dem zweiten Bewertenden für sämtliche Teilaspekte höher als für die Gruppe der Lehrkräfte bzw. den ersten schwedischen Bewertenden. Die durchgehend starken positiven Korrelationen zwischen den jeweiligen Bewerteraspekten in den GER-Urteilen und den globalen Bewertungen durch die schwedischen Bewertenden weisen zusammenfassend darauf hin, dass bei der Bewertung fremdsprachlicher

Schreibkompetenz ein Zusammenhang zwischen den unterschiedlichen Bewertungen besteht, auch wenn die Korrelationen mit den jeweiligen Aspektbewertungen leicht variieren.

8.4 Qualitativer Vergleich von Bewerterurteilen zweier grenzwertiger Leistungen

Bei einigen Textproduktionen auf *Tyska 5* sind sich die Bewertergruppen nicht einig, inwiefern die Texte das angestrebte Mindestniveau erreicht haben oder nicht, d. h. ob bei den schwedischen Bewertenden die Anforderungen für eine E-Note und bei den GER-Bewertenden die für ein B1-Niveau erfüllt sind. Da diese Mindestanforderungen in etwa in Relation zueinander stehen, erscheint es relevant, jegliche Bewerterurteile daraufhin zu untersuchen, inwiefern Unterschiede in dieser Hinsicht zu finden sind. Dies ist vor allem bei zwei Schülertexten auf *Tyska 5* der Fall. Es handelt sich hierbei um die beiden Schülerleistungen *Slsk1–5* und *Pnmj1–5*, die von zwei der schwedischen Bewertenden eine nicht ausreichende Note F erhalten haben, aber andererseits von einem schwedischen Bewertenden eine ausreichende Note (Note E), und die von den beiden GER-Bewertenden auf ein erreichtes B1-Niveau eingestuft wurden. Hier folgt ein qualitativer Vergleich zwischen den Bewerterkommentaren der Bewertergruppen zu diesen auffälligen und grenzwertigen Leistungen.

Die Schülerleistung ***Pnmj1–5: F/F/E*** hat von der praktizierenden Lehrkraft und der/dem ersten externen schwedischen Bewertenden eine nicht ausreichende Note F erhalten, wurde aber vom zweiten externen schwedischen Bewertenden mit der ausreichenden Note E bewertet. Alle schwedischen Bewertenden beschreiben sprachliche Mängel, vor allem im Bereich der *formalen Strukturen*, in der Schülerleistung. Sie sind sich aber nicht darüber einig, inwiefern diese sprachlichen Fehlgriffe auch das *Verständnis* beeinflussen. Während die Deutschlehrkraft in den Kommentaren angibt, dass das meiste verständlich ist, kommentiert die/der erste externe Bewertende, dass Teile der Leistung schwer zu verstehen sind. Die Bewertungen der schwedischen Bewertenden unterscheiden sich aber auch im Hinblick auf die inhaltliche *Aufgabenerfüllung* voneinander (vgl. Beispiele 8.1 und 8.2):

(8.1) Tre ganska korta texter som dock följer instruktionen.[178] (Pnmj1-5-E, ext. schwed. Bewert. 2)

(8.2) Uppgift 3 är ej fullföljd, då innehållet är ett annat än det efterfrågade.[179] (Pnmj1-5-F, Lehrkraft)

178 „Drei ziemlich kurze Texte, die aber den Instruktionen folgen“.

179 „Aufgabe 3 ist nicht gelöst, da sich der Inhalt von dem nachgefragten unterscheidet“.

Während die/der zweite externe Bewertende im Hinblick auf die Anforderungen in der Aufgabe einschätzt, dass die inhaltliche *Aufgabenerfüllung* angemessen ist (vgl. Beispiel 8.1), beschreiben dahingegen die/der erste externe Bewertende und die praktizierende Lehrkraft inhaltliche Mängel im Text (vgl. Beispiel 8.1). Sie kommentieren, dass der Inhalt unklar formuliert ist und von dem, was nachgefragt ist, abweicht. Die Lehrkraft schreibt in der Begründung zusätzlich, dass der Schüler/die Schülerin insgesamt im Fach Deutsch eine nicht ausreichende Note F als Abschlussnote riskiert. In der Gesamtheit kommentieren die schwedischen Bewertenden bei der Bewertung der Leistung *Pnmj1–5* die gleichen oder ähnlichen Aspekte in der Schülerleistung; sie scheinen aber die Erfüllung der inhaltlichen Anforderungen unterschiedlich zu interpretieren und außerdem scheinen sie sich nicht ganz einig zu sein, inwiefern die sprachlichen Mängel das Verständnis beeinträchtigen.

Die Textproduktion ***Pnmj1–5*** erhält von der/dem ersten und zweiten GER-Bewertenden insgesamt 70 bzw. 75 Punkte, was deutlich über der Bestehensgrenze von 60 Punkten liegt. Die GER-Bewertenden berücksichtigen Fehlgriffe im Bereich *Wortschatz* und *formaler Strukturen*. Sie schreiben aber in ihren Kommentaren, dass diese Fehlgriffe nicht oder nur stellenweise das *Verständnis* beeinträchtigen, z. B. gilt die Grußformel als nicht verständlich. Des Weiteren gilt die *Aufgabenerfüllung* gemäß den GER-Bewertenden als überwiegend angemessen und sie scheinen sich darüber einig zu sein. Die GER-Bewertenden beachten zudem auch Aspekte der *Angemessenheit* in der Schülerleistung. Sie kommentieren hierbei Kohärenz und Textaufbau generell in positiven Worten, wobei die/der zweite Bewertende aber auch einen Kommentar über einen Fehlgriff hinsichtlich der Textsorte abgibt. Außer einer Formulierung des ersten GER-Bewertenden über die Bewältigung einer Verbform (vgl. Beispiel 8.3) kommen in diesen Bewerterurteilen keine weiteren Hinweise darauf vor, dass dieser Text die Anforderungen eines B1-Niveaus nicht hätte erreichen sollen:

(8.3) der Passiv wird nicht beherrscht, ist aber Teil der Grammatik auf Niveau B1. (Pnmj1-5, GER-Bewert. 1)

Hier wird auf das Nicht-Beherrschen der Verbform im Passiv im Text verwiesen, was aber gemäß dem Bewertenden ein Teil der Grammatikkenntnisse auf dem Niveau B1 sein sollte. Zusammenfassend hat die Leistung in den GER-Bewertungen ein gut zufriedenstellendes Ergebnis bezüglich der berücksichtigten Aspekte erreicht und die GER-Bewertenden scheinen generell über die Bewertung ziemlich einig zu sein.

Auch die Textproduktion ***Slsk1–5: F/F/E*** hat voneinander abweichende Ergebnisse im Hinblick auf die Benotung nach schwedischen Bildungsstandards

erhalten. Wie bei der Textproduktion *Pnmj1–5* oben hat auch die Schülerleistung *Slsk1–5* von der praktizierenden Deutschlehrkraft und der/dem ersten externen Bewertenden eine nicht ausreichende Note F und vom zweiten externen Bewertenden eine ausreichende Note E erhalten. Alle drei schwedischen Bewertenden beschreiben sprachliche Mängel im Text, überwiegend im Bereich der *formalen Strukturen*. Diese Mängel beeinträchtigen allerdings nicht oder nur stellenweise das *Verständnis*. Sowohl die/der erste externe Bewertende als auch die praktizierende Lehrkraft betonen aber, dass gerade der Mangel an sprachlicher Präzision zu der Vergabe der Note F geführt haben (vgl. Beispiele 8.4 und 8.5):

(8.4) Går att följa trots språklig oprecision […] Då detta är tyska 5 borde precisionen för ett godkänt betyg vara bättre![180] (Slsk1-5-F, ext. schwed. Bewert. 1)

(8.5) Trots vissa kvaliteter bedöms inte texten motsvara kunskapskraven för E pga brister i språkets precision. Det är ju trots allt steg 5.[181] (Slsk1-5-F, Lehrkraft)

An den Beispielen ersichtlich scheinen die sprachlichen Mängel in der Schülerleistung für die Note entscheidend zu sein. Die/der zweite externe Bewertende, von dem die Leistung eine ausreichende Note E erhalten hat, kommentiert ebenfalls die Mängel, vor allem bezüglich der *formalen Strukturen*, gibt aber in der Begründung an, dass die Aufgabe trotzdem gelöst wird (vgl. Beispiel 8.6):

(8.6) Uppgiften genomförs och de efterfrågade delarna finns med.[182] (Slsk1-5-E, ext. schwed. Bewert. 2)

Im Hinblick auf die *Aufgabenerfüllung* wird von dem zweiten externen Bewertenden folglich argumentiert, dass die Textproduktion trotz gewisser sprachlicher Unklarheiten durchgeführt wird und die nachgefragten Teile enthält. Die Lehrkraft gibt an, dass der Lernende sich einfach ausdrückt und die inhaltlichen Anforderungen knapp erfüllt hat. Wie in den Bewerterkommentaren zum vorigen Textbeispiel, *Pnmj1–5*, können folglich Aspekte zur inhaltlichen Erfüllung der Aufgabe eine gewisse Rolle bei der divergierenden Benotung spielen. Zu bemerken ist aber, dass die inhaltliche Erfüllung hier nicht von allen schwedischen Bewertenden kommentiert wird. Es gibt auch weitere Aspekte, die nicht von sämtlichen Bewertenden in den Kommentaren berücksichtigt

---

180 „Man kann dem Text trotz unpräziser Sprache folgen […] Da dies Tyska 5 ist, sollte die Präzision für eine ausreichende Note besser sein“.

181 „Trotz gewisser Qualitäten wird nicht eingeschätzt, dass der Text aufgrund von einem Mangel sprachlicher Präzision die Wissensanforderungen für die Note E erfüllt. Es handelt sich hier immerhin um Stufe 5“.

182 „Die Aufgabe wird durchgeführt und die nachgefragten Teile sind enthalten“.

werden. Dies gilt für Aspekte zum *Textfluss*, zur *Kohärenz* und zur *soziokulturellen Angemessenheit.* Während z. B. die/der erste externe Bewertende auf eine fehlende Sie-Anrede in Anfangs- und Abschlussphrasen verweist, wird dies von den anderen beiden Bewertenden nicht kommentiert. Zusammenfassend scheinen die schwedischen Bewertenden teilweise unterschiedliche Begründungen zur Vergabe der Note zu geben und beachten dabei zum Teil auch verschiedene Aspekte.

Die Textproduktion ***Slsk1–5*** liegt mit 87 Punkten gemäß den beiden GER-Bewertenden deutlich auf einem erfüllten Niveau B1. In den Kommentaren werden u. a. Fehlgriffe im Bereich *Wortschatz* und *formaler Strukturen* beschrieben (vgl. Beispiel 8.7):

(8.7) Wortschatz. Mehrere Fehlgriffe beeinträchtigen das Verständnis nicht (Bitte, können die Jugend denken über andere Menschen auch). (Slsk1-5, GER-Bewert. 1)

Diese Fehlgriffe hinsichtlich *Wortschatz* und *formaler Strukturen* beeinträchtigen gemäß den beiden GER-Bewertenden jedoch nicht die *Verständlichkeit*. Auch wenn sprachliche Mängel vorkommen, scheinen sie dementsprechend für die Verständlichkeit der Leistung nicht entscheidend zu sein und führen nicht zu einer Einstufung unter dem B1-Niveau. Beide GER-Bewertende berücksichtigen zudem Aspekte der *Aufgabenerfüllung* und der *Angemessenheit*. Hierbei wird angegeben, dass die Anforderungen der Aufgabe inhaltlich und umfänglich angemessen behandelt sind. Ferner gilt die Lösung als *soziokulturell* angepasst und der *Textaufbau* als effektiv. Generell werden Aspekte der *Aufgabenerfüllung* und der *Angemessenheit* in sehr positiven Worten beschrieben und die GER-Bewertenden scheinen sich hierbei einig zu sein.

Zusammenfassend berücksichtigen die GER-Bewertenden generell bei der Einstufung in höherem Grad dieselben Aspekte und scheinen diesen Aspekten zudem ähnliche Bedeutung und gleiches Gewicht zuzumessen. Die schwedischen Bewertenden beachten dahingegen zum Teil unterschiedliche Aspekte und teilweise andere Aspekte als die, die in den GER-Bewertungen vorkommen, was den Vergleich zu den GER-Bewertenden schwieriger macht. Die schwedischen Bewertenden kommentieren häufig eher Aspekte, die in den Leistungen nicht erfüllt sind, während die GER-Bewertenden auch erfüllte Anforderungen bei den beachteten Aspekten beschreiben. Schwedische Bewertende zeigen, zumindest in diesem Vergleich, eine Tendenz, gelegentlich eine strengere Gewichtung sprachlicher Mängel, vor allem im Bereich der *formalen Strukturen* bzw. inhaltlicher *Aufgabenerfüllung*, vorzunehmen. Vor allem scheinen die Berücksichtigung unterschiedlicher Aspekte durch die Bewertenden und die

strengere Gewichtung gewisser Aspekte Gründe für die voneinander abweichenden Ergebnisse zwischen den schwedischen Bewertungen und den GER-Bewertungen zu sein.

8.5 Fazit

In diesem Kapitel wurde untersucht, in welcher Beziehung Lernproduktionen verschiedener Noten auf den jeweiligen Fremdsprachenstufen *Tyska 3*, *Tyska 4* und *Tyska 5* zu einem erreichten B1-Niveau stehen. Die Ergebnisse der vorliegenden Studie zeigen, dass eine knappe Mehrheit sämtlicher bewerteten Schülerleistungen in den Urteilen der GER-Bewertenden ein erreichtes GER-Niveau B1 erhalten haben. Die Anzahl der Schülerleistungen, die ein B1-Niveau erreicht, nimmt mit der Fremdsprachenstufe zu: 7/20 der Lernproduktionen auf *Tyska 3*, 11/20 derer auf *Tyska 4* und ganze 16/20 auf *Tyska 5*. Die Punktzahlen für Schülerleistungen auf den beiden Stufen *Tyska 3* und *Tyska* 4 sind im Durchschnitt jedoch unter der Bestehensgrenze für das Niveau B1. Schülerleistungen auf *Tyska 5* liegen durchschnittlich wesentlich höher und erhalten in höherem Grad das Prädikat *sehr gut* als Texte der beiden niedrigeren Stufen.

Die aktuellen Richtlinien für das Fach *Moderna Språk* nehmen die GER-Niveaus als Ausgangspunkt und geben dabei an, dass die Mindestanforderungen im Hinblick auf das sprachliche Kompetenzniveau der Lernenden nach dem Abschluss der fünften Fremdsprachenstufe, *Tyska 5*, sich an dem B1.2-Niveau des GER orientieren sollten. Auch wenn zu beachten ist, dass gerade die Testaufgaben der vorliegenden Arbeit im Unterricht nicht geübt oder vorbereitet wurden, ein Aspekt, den die eigenen Lehrkräfte womöglich bei der Bewertung berücksichtigen, kann festgestellt werden, dass bestandene Textproduktionen von schwedischen Schülerinnen und Schülern auf *Tyska 5* dem Erwartungsniveau B1 im Wesentlichen entsprechen. Den Ergebnissen nach ist das Verhältnis zwischen einem erreichten B1-Niveau und einer ausreichenden Note eindeutig: bis auf eine Ausnahme erreichen alle Schülertexte auf *Tyska 5*, die von mindestens einem schwedischen Bewertenden eine ausreichende Note E erhalten haben, gemäß den GER-Bewertungen das GER-Niveau B1 hinsichtlich der schriftlichen Kompetenz. Die große Mehrheit der Lernproduktionen auf *Tyska 5*, die die Anforderungen der schriftlichen Kompetenz nach schwedischen Kriterien erfüllen, scheint auf dem intendierten GER-Niveau zu liegen.

Dies bedeutet umgekehrt aber nicht, dass die Textproduktionen, die das B1-Niveau erreicht haben, unbedingt auch eine unterste Bestehensnote E erreicht haben. Das Verhältnis zwischen Textproduktionen auf *Tyska 5* mit einer nicht ausreichenden Benotung nach den schwedischen Kriterien und einem

erreichten B1-Niveau, ist somit weniger klar. Einige Schülerleistungen haben von schwedischen Bewertenden eine nicht ausreichende Note F erhalten, trotzdem aber bei der GER-Bewertung ein B1-Niveau erreicht. Die Tatsache, dass die schwedischen Bewertenden nicht über die F-Note übereinstimmen, indizierent, dass es sich hier um grenzwertige Leistungen handelt. Die große Mehrheit der Schülerleistungen, die von den schwedischen Bewertenden mit der Note F bewertet sind, erreichen aber nicht das B1-Niveau, Des Weiteren ist auch festzustellen, dass Schülerleistungen, die auf *Tyska* 3 oder *Tyska 4* die höheren Note A oder B erhalten haben, nach den GER-Bewerterurteilen ein erfülltes B1-Niveau erreichen.

Korrelationen zwischen den jeweiligen Bewertungen sowie zwischen einzelnen Teilaspekten bei den GER-Bewertungen und den entsprechenden schwedischen Bewertungen weisen relativ hohe Korrelationswerte auf, was relativ deutlich auf eine Beziehung hindeutet. Die Ergebnisse scheinen damit nahezulegen, dass ein ähnliches Konstrukt bewertet wird. Die Korrelationen zwischen den schwedischen Bewertungen und Teilaspekten bei der GER-Bewertung zeigen zudem, dass die Bewertungen der schwedischen Bewertenden in höherem Grad mit den Aspektbewertungen *Strukturen* und *Wortschatz* als mit der Aspektbewertung *Erfüllung* korrelieren.

Unter den beiden qualitativ untersuchten Textproduktionen, bei denen die GER-Bewertenden und die schwedischen Bewertenden bezüglich der Einstufung teilweise zu divergierende Ergebnissen gekommen sind, können gewisse Tendenzen wahrgenommen werden. Generell berücksichtigen die Bewertenden ähnliche Aspekte in den Schülerleistungen, auch wenn Aspekte wie z. B. *Aufgabenerfüllung* und *Angemessenheit* von den beiden GER-Bewertenden in etwas höherem Grad und positiver beachtet werden. Die schwedischen Bewertenden, die für die Leistungen die Note F vergeben haben, scheinen auch hinsichtlich der Anforderungen im Bereich *Wortschatz* und *formale Strukturen* und inhaltlichen *Aufgabenerfüllung* strenger, und in diesem Fall für die Note entscheidend, als die GER-Bewertenden zu bewerten.

Abschließend weisen sowohl die quantitative als auch die qualitativen Analysen auf ein starkes Verhältnis zwischen schwedischen Bewertungen der schriftlichen Sprachkompetenz auf den Stufen *Tyska 3*, *Tyska 4* und *Tyska 5* und den GER-Bewertungen hinsichtlich eines Niveaus B1 hin. Auch wenn die Bewertenden manchmal teilweise unterschiedlichen Aspekten bei der Einstufung Gewicht geben, kann die enge Beziehung wahrgenommen werden. Ferner zeigen die Ergebnisse generell, dass höhere Einstufungen im Hinblick auf

die Benotung im schwedischen System auch in höherem Grad ein erreichtes B1.2-Niveau bedeuten. Die Ergebnisse weisen zudem deutlich darauf hin, dass eine Beziehung zwischen den Mindestanforderungen auf *Tyska 5* und einem erreichten Sprachniveau B1 des GER hinsichtlich der schriftlichen Kompetenz vorliegt.

9. Diskussion

Der Fokus dieser Studie liegt auf Aspekten der Validität bei der Bewertung schriftlicher Sprachkompetenz im Fach *Tyska* am schwedischen Gymnasium. Betrachtet wurden hierbei relevante Aspekte der Validität bei der Bewertung einer Auswahl schriftlicher Lernproduktionen in diesem Fach im Hinblick auf a) die Konstruktkonzeptualisierung der Bewertenden hinsichtlich des zu messenden Konstrukts, b) die Bewerterübereinstimmung der schwedischen Bewertenden und c) die Beziehung schwedischer Bewertungen zu einem externen Referenzniveau B1 gemäß dem *Gemeinsamen europäischen Referenzrahmens für Sprachen* (GER). In diesem Kapitel folgt eine Diskussion und Interpretation der Ergebnisse dieser drei Teiluntersuchungen vor dem Hintergrund der Fragestellungen sowie aktueller Befunde der Forschung.

Anhand der Definition der Validität von Messick (1989b) soll durch ein integriertes Urteil ermittelt werden, *in welchem Grad* Inferenzen aus einem Testergebnis gezogen werden können. Hier bieten Kanes argumentbasiertes Validierungsmodell mit bestimmten Schritten hinsichtlich unterschiedlicher Inferenzen (vgl. Kane 2006; 2013 und hierzu auch Knoch & Chapelle 2018; Chapelle 2020) sowie verschiedene Aspekte der Validität innerhalb des soziokognitiven Rahmenmodells von Weir (2005, vgl. hierzu auch O'Sullivan & Weir 2011) einen guten Ausgangspunkt für die Diskussion. Zu bemerken ist, dass die Nachweise der Validität, die der vorliegenden Studie zugrunde liegen, nach dem Testerreignis, *a posteriori*, erhoben wurden. Somit werden hauptsächlich Aspekte der Validität in Betracht gezogen, die mit der Bewertung im Hinblick auf die Verwendung und Interpretation der Testergebnisse nach dem Testereignis zu tun haben. Da aber die verschiedenen Aspekte der Validität eng miteinander verbunden sind (vgl. Weir 2005), können auch Aspekte der Validität, die zum Testverlauf vor dem Testerreignis (*a priori*) gehören, für die Diskussion und Interpretation von Relevanz sein.

Basierend auf den Ergebnissen der vorliegenden Arbeit wird im Folgenden im Hinblick auf die Validität bei der Bewertung schriftlicher Sprachkompetenz die Konzeptualisierung des zu messenden Konstrukts unter Bewertenden reflektiert (Kap. 9.1). Es folgen Reflektionen zur Bewerterübereinstimmung zwischen den schwedischen Bewertenden (Kap. 9.2) und zur Beziehung zwischen schwedischen Bewertungen schriftlicher Kompetenz im Fach Deutsch am Gymnasium und GER-Bewertungen hinsichtlich des Sprachniveaus B1 (Kap. 9.3).

9.1 Inferenz der Bewertung und Begründung: Konstruktkonzeptualisierung der Bewertenden

Die Inferenz zur Bewertung (*scoring*) befasst sich mit der Frage, wie die Leistung eines Lernenden in ein beobachtetes Testergebnis umgesetzt wird. Hierbei soll also die beobachtete Leistung in ein beobachtetes Ergebnis umgewandelt werden, wobei angenommen wird, dass angemessene Bewertungskriterien verwendet werden (vgl. Kane 2013). Im Entscheidungsprozess ist die Konzeptualisierung der Bewertungskriterien unter den Bewertenden von großer Bedeutung. In einem erweiterten argumentbasierten Ansatz wird bei der Darstellung von Inferenzen zusätzlich auch eine Inferenz zur Begründung (*explanation*) miteinbezogen. Diese Inferenz bezieht sich u. a. darauf, inwiefern die Bewertungskriterien ein klar definiertes Konstrukt abdecken und inwiefern das Verständnis der Bewertenden mit dem zu messenden Konstrukt und mit den dahinterstehenden theoretischen Kompetenzmodellen konsistent ist (vgl. Knoch & Chapelle 2018). Nachweise der Validität im Hinblick auf die Inferenzen der Bewertung und Begründung, d. h. die Befunde zu den Bewerterkommenatern der Bewertenden bezogen auf die Schülerleistungen, werden in diesem Kapitel diskutiert.

Der Fokus dieser Studie liegt hauptsächlich auf der Perspektive der Bewertenden. Sie befasst sich hier mit der Frage, inwiefern Bewertende die gleichen oder unterschiedliche Aspekte auf der Ebene der Texte als relevant für die Bewertung wahrnehmen und inwieweit sich Unterschiede und Gemeinsamkeiten bezogen auf die Bewertergruppen oder unter den schwedischen Bewertenden finden lassen. Hier soll über die Ergebnisse zur Konzeptualisierung des Konstrukts von Bewertenden reflektiert werden, wobei verschiedene Aspekte der Konstruktvalidität, gemäß Messick (1989a) der zentrale Aspekt in Studien zur Validität, erörtert werden. Die Konstruktvalidität ist gemäß Weir (2005) als eine Funktion der Interaktion zwischen Aspekten der kognitiven Validität und Aspekten der Kontextvalidität in Verbindung mit den Bewertungskriterien zu verstehen. Im Folgenden wird die Untersuchung zum Bewertungsprozess im Hinblick darauf dargelegt, wie die Bewertenden Aspekte des zu messenden Konstrukts interpretieren können und bei der Bewertung einsetzen. Da kaum Studien zur Konzeptualisierung des Konstruktes aus einer Bewerterperspektive in einem schwedischen Schulkontext zu finden sind (eine Ausnahme ist Borger 2018), werden die Ergebnisse zudem mit internationalen Untersuchungen, die berücksichtigte Aspekte bei einer Bewertung fremdsprachlicher Kompetenz fokussiert haben (vgl. Kap. 4.1), verglichen und diskutiert.

Fokus der Bewertenden bei der Bewertung schriftlicher Kompetenz

Aus den qualitativen inhaltlichen Analysen der Bewerterkommentare ist zu entnehmen, dass die Bewertenden eine Vielfalt von unterschiedlichen Aspekten bei der Bewertung schriftlicher Kompetenz berücksichtigen. Dies zeigt generell auf eine breite Konzeptualisierung des Konstrukts unter den Bewertenden. Eine breite Konzeptualisierung des Konstrukts korrespondiert auch mit der breiten Darstellung von Komponenten in theoretischen Kompetenzstrukturmodellen kommunikativer Kompetenz (vgl. Kap. 3.1). Ein Ergebnis der empirischen Analysen ist folglich, dass insgesamt ein breites Spektrum von unterschiedlichen Aspekten in den Bewerterurteilen vorkommt, gleichzeitig ist aber nicht gesagt, dass die Bewertenden immer alle diese Aspekte in den einzelnen Urteilen berücksichtigten. Vielmehr kann stattdessen das breite Spektrum unterschiedlicher Aspekte in den Kommentaren auch auf individuelle Diskrepanzen unter den Bewertenden zurückgeführt werden, was auch bereits in anderen Studien festgestellt wurde (z. B. Eckes 2008; Kim 2009; Hsieh 2011; Borger 2018).

Hauptsächlich zeigen die Ergebnisse demgemäß, dass ein breites Spektrum unterschiedlicher Aspekte von den Bewertenden bei der Bewertung schriftlicher Kompetenz berücksichtigt werden, darunter vor allem solche wie *Angemessenheit*, *formale Strukturen*, *Wortschatz*, *Aufgabenerfüllun*g und *Verständlichkeit*. Die Tatsache, dass eine Vielfalt unterschiedlicher Aspekte beachtet werden, ist auch im Einklang mit aktuelleren Studien aus einem nordischen (vgl. Bøhn 2016; Borger 2018) und internationalen Kontext (vgl. Iwashita et al. 2008). In der vorliegenden Studie scheinen aber Aspekte zur *Verständlichkeit* im Vergleich zu bisherigen Studien häufiger in den Bewerterkommentaren vorzukommen. Sie werden bei Leistungen niedrigerer Niveaus häufiger kommentiert (vgl. Pollitt & Murray 1996) und dies könnte womöglich mit dem sprachlichen Niveau der Lernenden in dieser Studie zu tun haben. Die meistbeachteten Aspekte werden von beiden Bewertergruppen berücksichtigt, was generell auf ein ähnliches Verständnis des zu messende Konstrukts hindeutet. Die von den Bewertenden beachteten Aspekte und Bewertungsdimensionen sind zudem explizit oder zum Teil in den jeweiligen Bewertungskriterien, d. h. den schwedischen Bildungsstandards sowie den Skalen des GER, vertreten.

Weniger Kommentare gelten Aspekten, die nicht in den schwedischen Bildungsstandards vorkommen oder aus den Deskriptoren des GER stammen. Was womöglich als problematisch aufgefasst werden kann, ist die Tatsache, dass einige Aspekte aus dem Bewertungsraster (das den GER-Bewertenden zur Verfügung gestellt wurde) bzw. aus den Beurteilungsaspekten des schwedischen

nationalen Prüfungsmaterials zur Aufgabenerfüllung weder in den Deskriptoren des GER noch in den schwedischen Bildungsstandards zu finden sind. Diese Tatsache wird auch von Kecker (2011) hervorgehoben. Schwedische Bewertende, die das fakultative Bewertungsmaterial für *Moderna språk* kennen, können jedoch die enthaltenen Beurteilungsfaktoren zum Inhalt verwenden (vgl. Anhang 11).

Ferner kommentieren die schwedischen Bewertenden Aspekte wie *Gesamteindruck* und eine *pauschale Beurteilung der Sprache* bei der Bewertung fremdsprachlicher Leistungen, es wird ihnen aber in geringerem Ausmaß Aufmerksamkeit geschenkt. Zudem kommen Kommentare zu Aspekten, die als *kommunikative Strategien*, *Textfluss* oder *Sonstiges* einzuordnen sind, selten vor. Auffällig hierbei ist die geringe Anzahl von Segmenten, die in die Kategorien *Textfluss* und *kommunikative Strategien* eingeordnet werden können. Die Tatsache, dass Aspekte dieser Kategorien im untersuchten Datensatz im Vergleich zum Anteil der Kommentare in früheren Studien zur mündlichen Sprachkompetenz in wesentlich geringerem Ausmaß berücksichtigt werden (vgl. Brown et al. 2005, Iwashita et al. 2008; Hsieh 2011, Bøhn 2016; Borger 2018), könnte darauf hinweisen, dass die Flüssigkeit und die Verwendung kommunikativer Strategien wahrscheinlich von den Bewertenden eher bei der Bewertung mündlicher Sprachkompetenz beachtet werden. Auch wenn gerade Aspekte zur *Flüssigkeit* eher mit der mündlichen Sprachkompetenz verknüpft werden (siehe auch Europarat 2001: 129), werden *Flüssigkeit* und die Verwendung *kommunikativer Strategien* in den schwedischen Bildungsstandards mit sowohl schriftlichen als auch mündlichen Textproduktionen in Verbindung gesetzt (vgl. Kap. 2.2.3). Abschließend ist aber zu bemerken, dass viele der beachteten Aspekte sehr eng miteinander verbunden sind und daher auch schwer voneinander zu trennen sind. Dies zeigt wiederum, wie komplex eine Bewertung schriftlicher Kompetenz sein kann.

Zur Kategorie *Sonstiges* gehören Kommentare, die sich in die bereits erwähnten Bewertungskategorien nicht einordnen lassen. Hierzu zählen Kommentare verschiedener Ausdrucksweisen in den Textproduktionen (wie „mutig“ oder „humoristisch“), exakt übernommene Formulierungen aus den schwedischen Bildungsstandards (die zwar für die Bewertung schriftlicher Produktion relevant sind, jedoch nicht in diesem hier verwendeten Test geprüft werden), Hinweise auf übertragene Phrasen aus den Testaufgaben sowie Metakommentare über den Bewertungsprozess oder den Schwierigkeitsgrad der Aufgabe für die Lernenden. Die geringe Anzahl von Kommentaren, die als *Sonstiges* eingeordnet werden können, deutet allerdings darauf hin, dass die Bewertenden nur in geringem Ausmaß weitere Aspekte, die nicht zum zu messenden Konstrukt

gehören, berücksichtigen. Die Tatsache, dass kaum Nachweise von irrelevanten Variablen, die in die Bewertung miteinbezogen werden, sog. konstruktirrelevanter Varianz (vgl. Messick 1989b), vorliegen, ist von einem Qualitätsstandpunkt betrachtet vorteilhaft. Auch wenn die Ergebnisse der empirischen Analysen generelle Tendenzen hinsichtlich der Bewertung fremdsprachlicher Kompetenz aufzeigen, decken sie jedoch auch Unterschiede zwischen den Bewertergruppen auf, was im nächsten Abschnitt näher erläutert wird.

Unterschiede zwischen den Bewertergruppen

Auch wenn die beiden Bewertergruppen generell die gleichen oder ähnliche Aspekte beachten, deutet die qualitative Inhaltsanalyse der Bewerterkommentare auch auf Unterschiede zwischen den Bewertergruppen hin, u. a. im Hinblick auf die Anzahl der beachteten Aspekte pro Textproduktion. Die GER-Bewertenden verwenden hierbei ein Bewertungsmusterraster, das dazu einlädt, mehr zu kommentieren. Die GER-Bewertenden kommentieren dementsprechend pro Textproduktion mehr Aspekte als die schwedischen Bewertenden, wobei die schwedischen Bewertenden aber insgesamt, über die Texte verteilt, in höherem Ausmaß unterschiedliche Aspekte kommentieren. Dies könnte darauf hindeuten, dass die schwedischen Bewertenden in ihren Urteilen als Begründung für die Benotung unterschiedliche Aspekte anführen. Sie scheinen dementsprechend ihre Bewertungen teilweise auf unterschiedliche Aspekte zu gründen, da sie weniger Aspekte pro Textproduktion beachten. Nachweise dafür, dass die schwedischen Bewertenden in ihren Bewertungen zum Teil unterschiedliche Aspekte berücksichtigen und dieselben Aspekte unterschiedlich gewichten, sind zudem in den qualitativen Analysen einiger Bewerterurteile unterschiedlicher bzw. ähnlicher Benotungen zu finden (vgl. Kap. 7.4 bzw. Kap. 8.4).

Die Ergebnisse deuten folglich darauf hin, dass die Bewertergruppen, d. h. die schwedischen Bewertenden bzw. die GER-Bewertenden, sich im Hinblick darauf, welche Aspekte bei der Bewertung Berücksichtigung erhalten, zum Teil unterscheiden. Eine gewisse Variabililtät gehört aber zum Konstrukt, und die Unterschiede könnten auch auf kontextuelle Faktoren, wie die jeweiligen Bewerterskalen bzw. verschiedene Bewertungsverfahren, zurückgeführt werden. Vergleicht man die beiden Bewertergruppen, zeigt sich deutlich, dass sich diverse Unterschiede ergeben, die auf eine *analytische* bzw. *holistische* Herangehensweise bei der Bewertung zurückzuführen sind.

Die GER-Bewertenden beachten hauptsächlich Bewertungsdimensionen, die in irgendeiner Form im Bewertungsraster vorkommen. Es handelt sich hierbei

einerseits um die Aspekte *formale Strukturen* und *Wortschatz* sowie in gewissem Ausmaß deren Auswirkung auf die *Verständlichkeit* und andererseits um Aspekte zur *Erfüllung* der Anforderungen in der Aufgabe und verschiedenen Arten der *Angemessenheit.* Somit kann die Rangordnung der meistbeachteten Aspekte der GER-Bewertenden auf den großen Einfluss der eher analytisch ausgerichteten, zur Aufgabe gehörenden Bewertungsraster zurückgeführt werden. Die GER-Bewertenden scheinen zudem in höherem Ausmaß ihre Kommentare in positiven Worten zu verfassen. Ein analytisches Verfahren könnte hier dazu geführt haben, dass die Bewertenden bei der Beurteilung in höherem Grad Aspekte in den Schülerleistungen positiv einschätzen und dass versucht wird, die Qualitäten in den jeweiligen Texten zu finden.

Textproduktionen im schwedischen System werden generell mit einem holistischen Verfahren beurteilt. Wie auch im norwegischen System (vgl. Bøhn 2016) gibt es außer den Bildungsstandards für *Moderna språk* im schwedischen Schulkontext kein explizites Bewertungsraster für die Beurteilung schriftlicher Kompetenzen. Dahingegen ist im nationalen Prüfungsmaterial für die zweite Fremdsprache in Schweden, das u. a. für *Tyska 2, Tyska 3* und *Tyska 4* erhältlich ist, als Unterstützung für die Bewertung eine Darstellung eher analytischer Beurteilungsaspekte vorzufinden. Diese beachtet sowohl inhaltliche als auch sprachliche Aspekte (vgl. Anhang 11). Da aber das nationale Prüfungsmaterial für die zweite Fremdsprache lediglich fakultativ ist, ist unsicher, in welchem Ausmaß die Lehrkräfte diese Beurteilungsaspekte bei der Bewertung schriftlicher Produktion verwenden. Aus der Rangordnung der meistbeachteten Aspekte ist aber auch zu entnehmen, dass *Aspekte der linguistischen Kompetenz* generell unter den schwedischen Bewertenden eine etwas größere Rolle zu spielen scheinen. Unterschiede zwischen den Bewertergruppen ergeben sich aber, wenn die Kategorien *pauschale Bewertung der Sprache* und *Gesamteindruck* betrachtet werden. Die Tatsache, dass die schwedischen Bewertenden relativ häufig einen globalen Eindruck sowohl im Hinblick auf die pauschale Beurteilung der Sprache als auch auf den gesamten Text formulieren, könnte mit der eher holistischen Herangehensweise bei der Bewertung in Verbindung gesetzt werden. Allerdings hätten auch die GER-Bewertenden im Anschluss an die eher analytische Bewertung eine Art Gesamteindruck geben können, was hier jedoch nicht verlangt wurde.

Sowohl die GER-Bewertenden als auch die schwedischen Bewertenden beachten Aspekte zur *Angemessenheit*. Insgesamt fällt die relativ große Anzahl der Kommentare zur *soziokulturellen Angemessenheit* unter den schwedischen Bewertenden im Vergleich zu anderen Aspekten auf. Hierbei scheinen für die schwedischen Bewertenden ein angemessenes Verwenden von *Sie* oder *du* sowie

partneradäquate Grußformeln in den Textproduktionen im Hinblick auf soziokulturelle Angemessenheit bei der Bewertung von Bedeutung zu sein. In den schwedischen Bildungsstandards wird eine situations- und partneradäquate Anpassung verlangt, die u. a. die soziolinguistische Angemessenheit abdeckt (vgl. Kap. 2.2.3). Insbesondere auf *Tyska 5* sollen die Lernenden in formelleren Kontexten ein formelles Register verwenden können. Hinweise auf die übrigen Teildimensionen, wie Aspekte zur *Kohärenz*, zur *Textsorte* und zum *Textaufbau*, sind jedoch auch, mehr oder weniger explizit, in den Formulierungen der schwedischen Bildungsstandards zu finden. Ein Grund für die Hervorhebung *soziokultureller Angemessenheit* bei den schwedischen Bewertenden könnte eventuell die Aufgabenstellung des Tests sein, da in den beiden E-Mails unterschiedliche Grade der Formalität verlangt werden.

Bei den beiden GER-Bewertenden überwiegen dahingegen, trotz der expliziten Erwähnung im Bewertungsraster, Kommentare zur *Kohärenz* und zum *Textaufbau* über Kommentare zur *soziokulturellen Angemessenheit* und *Textsorte*, was allerdings zeigt, dass Faktoren wie die Aufgabenstellung und eine Erwähnung im Bewertungsraster allein diese Unterschiede nicht erklären kann. Sowohl im Bewertungsraster der GER-Bewertenden (vgl. Anhang 12) als auch in den Beschreibungen der soziolinguistischen bzw. pragmatischen Kompetenzen des GER (vgl. Europarat 2001: 118 ff.) wird auf die jeweiligen Teildimensionen, wie *Register/soziokulturelle Angemessenheit*, *Textgestaltung/Textsorte*, *thematische Organisation/Textaufbau* und *Kohärenz/Kohäsion,* hingewiesen.

Weitere Unterschiede finden sich unter den Aspekten, die sich auf den Inhalt in den Textproduktionen beziehen. Kommentare zu Aspekten hinsichtlich der *Aufgabenerfüllung* kommen in den Urteilen der schwedischen Bewertenden weniger vor. Die GER-Bewertenden kommentieren dagegen in ziemlich hohen Grad sowohl die Textlänge als auch die inhaltliche Aufgabenerfüllung. Ein Grund für die Unterschiede zwischen den Bewertergruppen im Hinblick auf diese Aspekte könnte womöglich in der schwedischen Lerntradition liegen. In den Anweisungen oder Aufgaben des schwedischen nationalen Testmaterials wird häufig nicht explizit angegeben, aus welchen Elementen eine Schülerleistung bestehen sollte oder auf eine bestimmte Wortanzahl hingewiesen. Die Subkategorie *Textlänge* ist in diesem Zusammenhang aus dem Grund interessant, dass die Wortanzahl im hier verwendeten Test angegeben wird. Die Textmenge wird aber normalerweise in einem schwedischen Schulkontext nicht angegeben, z. B. auch nicht im nationalen Prüfungsmaterial für die Fremdsprachen. Vielmehr sind die Aufgaben des schriftlichen Ausdrucks in einem schwedischen Schulkontext darauf ausgerichtet, dass die Lernenden nachweisen sollen, dass sie über genügende Deutschkenntnisse

verfügen, um eine realitätsnahe Situation bewältigen zu können. Dabei sind folglich häufig keine genaueren Vorgaben angegeben und die Gestaltung des Inhaltes ist mehr oder weniger den Schülerinnen und Schülern überlassen. Diese offene Struktur im Hinblick auf die inhaltliche Aufgabenerfüllung und die Textlänge könnte zu unterschiedlichen Interpretationen und Gewichtungen unter den Lehrkräften geführt haben. Aufgabenspezifische Bewertungskriterien, die konkret und deutlich inhaltliche Aspekte behandeln, sind daher bei der Bewertung nicht zu unterschätzen, um ein gemeinsames Verständnis diesbezüglich herzustellen.

Des Weiteren enthalten die Bewerterurteile der schwedischen Bewertenden im Gegensatz zu den Urteilen der GER-Bewertenden Aspekte zu *kommunikativen Strategien*, wenn auch in relativ geringem Ausmaß. Aspekte zu *kommunikativen Strategien* scheinen allgemein etwas häufiger bei der Bewertung mündlicher Sprachkompetenz beachtet zu werden (vgl. Borger 2018). Die Verwendung kommunikativer Strategien, um sprachliche Schwierigkeiten zu lösen, z. B. durch Umformulierungen oder Erklärungen, ist in den schwedischen Bildungsstandards zu finden. Es kann manchmal aber schwierig sein, das Verwenden kommunikativer Strategien in schriftlichen Lernproduktionen wahrzunehmen, und dies könnte hier ein Grund für die relativ geringe Anzahl von Kommentaren zu dieser Kategorie sein. *Kommunikative Strategien*, wenn von schwedischen Bewertenden mithilfe eines analytischen Bewertungsrasters kommentiert, scheinen manchmal eher als eine nachträgliche Rechtfertigung für die Benotung zu funktionieren (vgl. Lumley 2002): Diese Kommentare beziehen sich zudem eher auf die Formulierung in den Bildungsstandards als auf Aspekte in den zu bewertenden Textproduktionen.

Zusammenfassend scheinen somit kontextuelle Faktoren, wie die Bewertungsskalen und das Bewertungsverfahren, für die Bewertung schriftlicher Kompetenz von Bedeutung zu sein. Die Konzeptualisierung für das zu messende Konstrukt scheint aber sich auch zwischen den Bewertergruppen zu unterscheiden. Einiges spricht hierfür, die GER-Bewertungen beachten u. a. in größerem Ausmaß, inwiefern Aspekte die linguistische Kompetenz das Verständnis beeinträchtigen oder nicht. Unter den schwedischen Bewertenden ist dieses Verhältnis hingegen im Hinblick auf das zu messende Konstrukt weniger deutlich.

Es kann hiermit nicht verneint werden, dass das Bewerterverhalten eine große Rolle zu spielen scheint. Die Ergebnisse weisen darauf hin, dass schwedische Bewertende mit dem holistischen Verfahren ein vergleichsweise breiteres Spektrum von Aspekten berücksichtigen. Wiederum weist die Rangordnung der beachteten Aspekte der GER-Bewertenden auf einen großen Einfluss des

Bewertungsverfahrens für Auswahl und Verteilung der in den Urteilen berücksichtigten Aspekte hin: Die meistberücksichtigten Aspekte sind in irgendeiner Form im analytisch eingerichteten Bewertungsraster, das die GER-Bewertenden verwendet haben und das zum schriftlichen Test gehört, vertreten. Der große Einfluss des Bewertungsverfahrens zeigt sich auch bei den wenigen Lehrkräften, die ausschließlich oder ergänzend ein analytisches Bewertungsraster verwenden. Diese lokal verwendeten Bewertungsraster basieren auf den schwedischen Bildungsstandards, enthalten aber trotzdem nicht immer die gleichen Bewertungsdimensionen: Während einige Raster Aspekte wie *kommunikative Strategien* mit in Betracht ziehen, enthalten andere stattdessen Aspekte wie den *Textfluss* oder die *soziokulturelle Angemessenheit*. Darüber hinaus scheint es der Fall zu sein, dass ein Bewertungsraster dazu einladen kann, mehr zu kommentieren.

Dies zeigt insgesamt, dass das Bewertungsverfahren einen Einfluss darauf haben kann, auf *welche* Aspekte Bewertende bei der Bewertung ihre Aufmerksamkeit richten und darauf, wie sie diese verstehen, gewichten und interpretieren. Auch wenn die Bedeutung des Bewertungsverfahrens in der vorliegenden Studie nicht im Zentrum steht, kann abschließend festgehalten werden, dass das jeweilige Bewertungsverfahren einen großen Einfluss auf die Bewertung haben kann. Bei einer analytischen bzw. holistischen Bewertung sollte dies berücksichtigt werden, damit die Risiken der jeweiligen Verfahren (vgl. Crooks et al. 1996; Weigle 2002) nicht vernachlässigt werden und das Verfahren für die Interpretation und Verwendung von Testergebnissen keine allzu große Bedeutung erhält. Darüber hinaus konnte in früheren Studien festgestellt werden, dass auch Hintergrundsfaktoren der Bewertenden, wie ein gemeinsamer oder ähnlicher Ausbildungshintergrund und Berufserfahrung, Bewertungen in dieselbe Richtung beeinflussen (vgl. Song & Caruso 1996; Cumming et al. 2002): so scheinen in der vorliegenden Studie die GER-Bewertenden, die gleiche Ausbildung absolviert haben und deren Bewertungen kontinuierlich vom Sprachinstitut kontrolliert werden, eine gemeinsame Basis für die Bewertung schriftlicher Kompetenz zu haben.

Unterschiede zwischen den schwedischen Bewertenden

Auch wenn die vorliegende Arbeit die Variation innerhalb der Bewertergruppen nicht speziell untersucht, ist deutlich, dass die empirischen Ergebnisse auf gewisse Unterschiede zwischen den jeweiligen schwedischen Bewertenden hinweisen. Die Tatsache, dass einzelne Bewertende individuelle Schwerpunkte haben oder Aspekte bei der Bewertung unterschiedlich gewichten ist zwar

nicht überraschend und könnte u. a. mit der Berufserfahrung, Ausbildung oder individuellen Profilen der Bewertenden zu tun haben.

Eine besondere Aufmerksamkeit erhalten jedoch sprachliche Korrekturen bezüglich Grammatik, Orthographie sowie Wortschatz von der Gruppe der schwedischen Lehrkräfte. Sie achten dabei vergleichsweise weniger auf inhaltliche Anforderungen. Diese Gruppe besteht allerdings aus 18 schwedischen Lehrkräften und es ist daher anzunehmen, dass es eine bedeutende Variation innerhalb diese Gruppe gibt und dass der Form-Fokus bei den schwedischen Lehrkräften nicht von allen ausgeht. Dies hat womöglich eine Auswirkung auf die Ergebnisse, insgesamt kann man jedoch von einer allgemeinen Tendenz sprechen. Hierbei können Erfahrung und Hintergrund der Lehrkräfte möglicherweise eine Rolle spielen und die Profile der Lehrkräfte beeinflussen.

Diese Tendenz ist im Einklang mit vorherigen Untersuchungen, die gezeigt haben, dass gerade praktizierende Lehrkräfte häufig ihre Aufmerksamkeit auf Mängel hinsichtlich Orthographie und sprachlicher Korrektheit richten (vgl. Birkel & Birkel 2002; Kuiken & Vedder 2014). Die Tatsache, dass ein gewisser Fokus auf der sprachlichen Form zu liegen scheint, könnte womöglich mit einer Unterrichtstradition hinsichtlich des Fremdsprachenlernens zusammenhängen. Es könnte sich hier dementsprechend um eine Frage der Priorität handeln. Der Deutschunterricht im schwedischen Schulkontext hat eine lange Tradition, sich mit der Form zu beschäftigen (vgl. SOU 1948:27), was insbesondere für formreiche Sprachen wie Deutsch und Französisch der Fall zu sein scheint (vgl. Tornberg 2000). Trotz des heutigen Fokus auf einen handlungsorientierten Ansatz könnte es sein, dass wir uns immer noch teilweise in der Spannung zwischen diesen beiden Polen befinden.

Eine Auswahl von bestimmten Aspekten bei der Bewertung, in diesem Fall die Betonung sprachlicher Korrektheit auf Kosten der Aufgabenerfüllung, könnte möglicherweise zu Konsequenzen hinsichtlich der Validität führen, in diesem Fall eine Gefahr der Unterrepräsentation des Konstrukts. Hierbei besteht die Gefahr, dass andere Bewerteraspekte nicht genügend beachtet werden oder dass die sprachliche Korrektheit die anderen zu bewertenden Aspekte beeinflusst, sog. *Halo-Effekte*. Halo-Effekte können beispielsweise vorkommen, wenn die Bewertungsentscheidung zu schnell gefallen ist (vgl. Bortz & Döring 2002). Inwiefern dies hier tatsächlich der Fall sein könnte, müsste aber durch andere Studien bestätigt werden.

Des Weiteren scheint der Unterrichtskontext einen Einfluss auf die Bewertung zu haben. Der mögliche Fokus auf sprachliche Korrekturen könnte eventuell auch damit zu tun haben, dass die Lehrkräfte wissen, was sie im Unterricht

behandelt haben oder was nach ihren Vorstellungen von den Schülerinnen und Schülern verschiedenen Niveaus zu erwarten ist vgl. (Jølle 2015). Hierbei wollen sie prüfen, inwieweit die Lernenden die besprochenen, häufig grammatischen, Phänomene gelernt haben. Die praktizierenden Lehrkräfte richten gelegentlich zudem die Kommentare direkt an ihre jeweiligen Schülerinnen und Schüler. Wenn an Lernende gerichtet, drücken die Lehrkräfte sich vielleicht anders aus als wenn sie eine Begründung schreiben, die ausschließlich von anderen Lehrkräften oder Forschern gelesen werden soll.

Darüber hinaus kann eine Diskussion über Kommentare geführt werden, in denen einige der schwedischen Lehrkräfte ausschließlich die Mindestanforderungen für die jeweilige Notenstufe in den schwedischen Bildungsstandards exakt zitiert haben. Diese Kommentare der Lehrkräfte sind problematisch, da nicht deutlich wird, inwiefern die Lehrkräfte auch tatsächlich diese Aspekte bei der Bewertung beachtet haben. Es könnte sich hier um einen Ausdruck für eine Unsicherheit der Lehrkräfte darüber handeln, welche Aspekte sie in den Schülertexten bei der Bewertung beachten sollten. Durch ein Zitieren der Kriterien der jeweiligen Notenstufen für die Fremdsprachen kann eine Stellungnahme dazu, welche Aspekte oder Dimensionen bei der Bewertung jener Textproduktion wahrgenommen wurden, vermieden werden.

Zum Teil deuten die Ergebnisse der schwedischen Bewertungen darauf hin, dass die Bewertenden Aspekte unterschiedlich interpretieren und gewichten. Die Befunde der schwedischen Bewerterurteile deuten darauf hin, dass schwedische Bewertende zum Teil unterschiedliche Interpretationen des Konstruktes haben. Wenn Bewertende unterschiedliche Meinungen hinsichtlich des zu messenden Konstruktes haben, ruft das häufig Kritik gegen vage Bewertungskriterien hervor (z. B. Wisniewski 2010). Inwiefern offen gehaltene Bewertungskriterien immer von Nachteil sein müssen, kann jedoch diskutiert werden. Stobart (2012) sieht jedoch eine Gefahr, wenn Kriterien zu dem Grad detailliert werden, „that they encourage impoverished learning“ (S. 238). Detaillierte Kriterien im Hinblick darauf, wie bestimmte Note oder Niveaus zu erreichen sind, könnten die Herausforderung am Lernen in Frage stellen und somit die Validität der Leistungen verringen. Torrance (2007) beschreibt dies als eine Bewegung „from assessment *of* learning, trough assessment *for* learning, to assessment *as* learning“ (S. 281, *Hervorheb. im Original*). Eine intendierte Klarheit und Explizitheit beim Formulieren von Bewertungskriterien könnten folglich zu einer allzu großen Detailliertheit bei der Bewertung führen, was wiederum einer Fragmentarisierung des zu messenden Konstrukts mit sich bringen könnte. Hierbei ist es wichtig zu beachten, was auch Erickson (2020a) betont, dass dies nämlich nicht zu einer Verengung des zu messenden

Konstruktes führt, wodurch die Gefahr entstünde, dass eher das beurteilt wird, was leicht überprüfbar ist (vgl. Erickson & Åberg-Bengtsson 2012), und nicht die authentische Sprachkompetenz. Ein allzu großer Fokus auf Bewertungsverfahren könnte überdies zur Folge haben, dass die Lernerfahrung davon überragt wird und dass dabei andere Werte, wie intellektuelle und soziale Anregungen, nicht beachtet werden.

Das abschließende Bild ergibt, dass die hier untersuchten Bewerterurteile viele Gemeinsamkeiten haben, aber es lässt sich auch feststellen, dass Unterschiede zu finden sind. Es stellt sich die Frage, inwieweit diese Unterschiede auf die Formulierungen der jeweiligen *a priori* vergebenen Bewertungskriterien zurückzuführen sind, d. h. die schwedischen Bildungsstandards für die zweite Fremdsprache bzw. die GER-Skalen oder das Bewertungsraster zum Test, oder auf andere Faktoren, wie kontextgebundene Unterschiede oder tradierte Bewertungstraditionen im Fach Deutsch. Gemäß Jølle (2015) haben unterschiedliche Schulfächer verschiedene Traditionen, aber auch wenn die Lehrkräfte die jeweiligen Bewertungskriterien berücksichtigen müssen, gibt es andere Aspekte als das, was in den Kriterien zum Ausdruck kommt, die sie in ihren Bewertungen miteinbeziehen sollen. Er beschreibt die Spannung zwischen expliziten Kriterien und Kriterien, die als gegeben anzunehmen sind. Es handelt sich hierbei nach Jølle nicht nur um unterschiedliche Interpretationen, sondern auch um *rater values* und *rater choices* – und hier zeigt sich erneut, wie herausfordernd und komplex die Bewertungspraktiken bei der Einstufung schriftlicher Leistungen sein können.

Hinzu kommt auch, dass die schwedischen Bewertenden zu unterschiedlichen Zeitpunkten ihre Lehrerausbildung absolviert haben und danach in unterschiedlichem Grad Fortbildungen oder Kurse im Bereich Bewertung belegt haben. Des Weiteren könnte angenommen werden, dass erfahrene Lehrkräfte, die zusätzlich als Bewertende tätig waren, häufiger in Kontakt mit kommunikativen Strömungen hinsichtlich Testen und Bewertung gekommen sind und somit ein breiteres Bild der Sprachkompetenz auch bei der Erfüllung der Aufgabe beachten. Dies kann mit der Prüferschulung der GER-Bewertenden in Verbindung gesetzt werden, die als Voraussetzung für Prüfende des Goethe-Zertifikats verlangt wird und regelmäßig erneut werden muss. Inwieweit Ausbildungshintergrund, kontextgebundene Unterschiede oder tradierte Bewertungstraditionen letztendlich einen Einfluss auf die Bewertung schriftlicher Kompetenz in Deutsch haben, lässt sich hier aber schwer abschließend beantworten und die Ergebnisse sollte daher mit anderen Studien ergänzt werden.

9.2 Inferenz der Generalisierung: Aspekte der Validität bei der Ergebnisermittlung

Die Inferenz zur Generalisierung (*generalization*) befasst sich u. a. mit der Übereinstimmung von Bewertungen (vgl. Kane 2006, 2013). Im Hinblick darauf wird in der vorliegenden Arbeit hinsichtlich der zweiten Forschungsfrage untersucht, inwiefern die Testergebnisse der Bewertungen durch die schwedischen Bewertenden übereinstimmend sind und inwieweit dabei Nachweise für Reliabilität bezüglich der Testergebnisse gezeigt werden können. Fragen hinsichtlich der Reliabilität werden nach Messicks einheitlichem Validitätskonzept als Teil der Nachweise für die Testinterpretation angesehen (vgl. hierzu Kap. 3.2, Tab. 9) und die Reliabilität wird zudem in Weirs sozio-kognitivem Rahmenmodell (2005) unter dem Begriff Validität der Ergebnisermittlung als wichtiger Teil der Validität angesehen. Zu den Aspekten bezüglich der Validität der Ergebnisermittlung, gemäß Weir auch eng mit Aspekten zur Kontextvalidität und zur kognitiven Validität verbunden, gehört die Bewerterübereinstimmung unter Bewertenden (2005: 24).

Nicht nur in der Forschung sind Nachweise der Reliabilität von Bedeutung; auch im schwedischen Schulkontext wird Aufmerksamkeit auf Fragen zur Reliabilität gerichtet, insbesondere hinsichtlich der Bewerterübereinstimmung (vgl. Kap. 4.2). In der vorliegenden Arbeit sind unterschiedliche Methoden zur Ermittlung der Beurteilerübereinstimmung zwischen den jeweiligen Bewertenden verwendet worden. Diese Methoden beruhen auf unterschiedlichen Annahmen und daher können aus den Ergebnissen unterschiedliche Typen von Schlüssen gezogen werden. Da der schwedische Schulkontext den Ausgangspunkt der vorliegenden Studie bildet, werden die Ergebnisse hauptsächlich mit Studien verglichen und diskutiert, die Bewertungen im Hinblick auf die Bewerterübereinstimmung innerhalb des schwedischen Systems untersucht haben.

Bewerterübereinstimmung: Konsens und Konsistenz

Die Befunde dieser Studie zeigen, wenn gängige Reliabilitätswerte ermittelt werden, nicht immer zufriedenstellende Resultate zwischen den verschiedenen schwedischen Bewertungen. Hierbei kann aber ein Unterschied bezüglich der Ermittlungen der Konsens- bzw. Konsistenzwerte beobachtet werden. Die Ermittlungen zum Konsens zwischen den Bewertenden, die ein Hinweis darauf sind, in welchem Ausmaß die Bewertenden zum genau gleichen Ergebnis gekommen sind, zeigen hier niedrigere Werte. Dies scheint insbesondere dann

der Fall zu sein, wenn die Ergebnisse der jeweiligen externen Bewertenden mit den Bewertungen aus der Gruppe der Lehrkräfte verglichen werden. Die Konsenswerte in der vorliegenden Studie liegen zum Teil unter einem akzeptablen Niveau, was beachtet werden sollte. Konsenswerte können für Unterschiede hinsichtlich der Tendenz zur Milde/Strenge empfindlich sein, aber auch für Variabilität in der Notengebung. Sie könnten somit einen Hinweis darauf geben, dass Bewertende unterschiedlich streng bewerten sowie dass sie unterschiedliche Interpretationen vornehmen und Kriterien unterschiedlich gewichten. Konsenswerte unterhalb eines akzeptablen Niveaus sind jedoch bei Bewertungen freier Textproduktion auch in anderen Studien zu finden (z. B. Eckes 2011; Tengberg et al. 2017). Es ist dementsprechend nicht ungewöhnlich, dass Bewertende nicht bei allen Bewertungen freier Produktion Konsens erreichen.

Die Ermittlungen zur Bestimmung der Konsistenz weisen dahingegen darauf hin, dass die schwedischen Bewertenden generell bei den Bewertungen relativ hohe und zufriedenstellende Konsistenzwerte aufweisen. Die Ergebnisse der Konsistenz sind vergleichsweise höher (vor allem hinsichtlich der Rangkorrelationen) oder im Einklang mit bisherigen Studien aus dem schwedischen Schulkontext, die Bewertungen freier Produktion untersucht haben (z. B. Erickson 2009; Borger 2018). Diese Befunde deuten darauf hin, dass die Differenzen unter den schwedischen Bewertenden bei der Rangfolge der Leistungen dementsprechend nicht so groß sind, aber sehr wohl bei der Notengebung.

Der Unterschied zwischen den Ergebnissen hinsichtlich der Konsens- bzw. Konsistenzwerte in der vorliegenden Studie weist somit darauf hin, dass die exakte Übereinstimmung zwischen den schwedischen Bewertenden bei der Bewertung niedrigerer ausfällt, aber dass ihre Bewertungen in der gleichen Relation zueinander stehen. Die Bewertenden scheinen demzufolge unterschiedliche Noten zu vergeben, sie können aber die Fremdsprachenkenntnisse der Lernenden gut einschätzen und sind in höherem Grad darin übereinstimmig, welche Leistungen in Relation zu den Kriterien besser oder schlechter ausfallen. Das Ergebnis, dass Unterschiede zwischen Konsens- bzw. Konsistenzwerten zu finden sind, steht im Einklang mit anderen bisherigen Studien aus dem schwedischen Schulkontext (z. B. SOU 1942:11; Johansson 2013; Tengberg et al. 2017) sowie mit Studien zur Bewertung schriftlicher Produktionen in Deutsch als Fremdsprache (z. B. Bärenfänger 2016), die in ihren Analysen das gleiche Verhältnis gefunden haben.

Wie und bei welchen Notenstufen entstehen häufiger divergierende Urteile bei der Bewertung? Durch die Kreuztabellen kann festgestellt werden, dass die Bewertenden gerade bei der Bewertung niedrigerer Noten in höherem Grad

übereinstimmen und dass die Bewertung von Textproduktionen im mittleren und oberen Bereich des Notensystems häufiger zu unterschiedlichen Noten führt. Vor allem zeigen die Ergebnisse der vorliegenden Studie, dass die Bewertenden bei der Bewertung einer nicht ausreichenden Leistung (Note F) etwas häufiger miteinander übereinstimmen. Dies ist insbesondere interessant, da Lehrkräfte es nach eigenen Angaben manchmal problematisch finden und ein Gefühl der Unsicherheit im Hinblick darauf haben, was die Schülerinnen und Schüler mindestens leisten müssen, um ein ausreichendes Niveau zu erreichen (vgl. Erickson 2009; Papageorgiou 2010; Håkansson Ramberg 2021). Auch wenn Lehrkräfte vor allem bei der Benotung zwischen nicht-erreichten und knapp-erreichten Noten zu überlegen scheinen und gerade diese Texte untereinander diskutieren, ist es dementsprechend sinnvoll, Lehrkräfte auch zur Bewertungsdiskussion von Texten mittleren und höheren Niveaus zu ermuntern. Es scheint offenbar einfacher, wie auch in anderen Studien gezeigt (z. B. Hambleton et al. 1995; Birkel & Birkel 2002; Erickson 2009; Granfeldt & Ågren 2014; Skolinspektionen 2017), unter weniger avancierten Leistungen eine Übereinstimmung zu finden als zwischen Textproduktionen im mittleren oder höheren Notenbereich. Gemäß den qualitativen Analysen der Bewerterkommentare berücksichtigen die schwedischen Bewertenden bei Schülerleistungen divergierender Benotung häufig die gleichen Dimensionen, gewichten sie aber unterschiedlich. Im Fall der nicht ausreichenden Note F haben sie dahingegen in höherem Ausmaß dieselben Maßstäbe für Mindestleistungen und kommen daher die Bewertenden gerade bei diesen Textproduktionen häufiger zu demselben Urteil. Es scheint demnach einfacher zu bestimmen, wann die Anforderungen nicht erfüllt sind, als den Grad der Erfüllung einzuschätzen. Eine mögliche Erklärung bei einigen Bewertungen ist, dass eine der drei Aufgaben fehlt. Dies führt eher automatisch zu einer nicht-erreichten Note und es ist daher einfacher zu einer Übereinstimmung zu kommen.

Tendenzen zur Milde und Strenge

Die Gruppe der Lehrkräfte in der vorliegenden Studie hat die Textproduktionen ihrer eigenen Schülerinnen und Schüler bewertet. Sie sollten dabei ihrem normalen Bewertungsablauf folgen. Die Schülerleistungen waren für sie deswegen nicht anonymisiert und sie hatten dementsprechend Zugang zu den Namen der Probanden. Es muss in diesem Zusammenhang aber nochmals erwähnt werden, dass die Gruppe der schwedischen Lehrkräfte aus unterschiedlichen Individuen besteht und ihre Teilnahme zudem im Material in unterschiedlichem Ausmaß repräsentiert ist. Die zum Teil ziemlich bedeutenden

Notenabweichungen könnten dementsprechend auf gewisse individuelle Lehrkräfte zurückgehen. Allerdings kann, wie in anderen Studien (z. B. Harlen 2005; Skolinspektionen 2018), in der vorliegenden Arbeit beobachtet werden, dass die Gruppe der Lehrkräfte im Vergleich zu den externen Bewertenden im Durchschnitt den Textproduktionen höhere Noten gibt. Dies zeigt sich nicht zuletzt durch die Multifacetten-Rasch-Analyse, aber auch in der deskriptiven Statistik im Hinblick auf die Mittelwerte der jeweiligen schwedischen Bewertenden und die Verteilung der Bewertungen pro Bewertenden über die jeweiligen Notenstufen. Die Gruppe der Lehrkräfte hat über alle Fremdsprachenstufen einen deutlich höheren Notendurchschnitt (vgl. Tab. 29) und es geht aus der Multifacetten-Rasch-Analyse (Abb. 10) klar hervor, dass diese Gruppe im Vergleich zu den beiden externen Bewertenden etwas milder bewertet.

Wiederum zeigen die externen schwedischen Bewertenden im Vergleich zu der Gruppe der Deutschlehrkräfte eine leichte Tendenz zur Strenge. Die Tendenz, eine nicht ausreichende Note zu vergeben, ist insbesondere deutlich, wenn die Notenverteilung des ersten externen Bewertenden betrachtet wird (vgl. Abb. 9). Insgesamt etwa ein Drittel der Textproduktionen hat von der/dem ersten externen Bewertenden eine nicht ausreichende Note erhalten. Da in bisherigen Studien geklärt wurde, dass gerade externe Bewertende durch ihre externe Bewerterposition eine Tendenz zu Strenge haben könnten (vgl. Gustafsson & Erickson 2013), sollten die Ergebnisse der externen Bewertenden aus diesem Grund mit Vorsicht interpretiert werden. Anderseits haben die externen Bewertenden der vorliegenden Studie umfangreiche Erfahrungen im Bereich Bewertung von Schülertexten und die Beurteilerübereinstimmung dieser beiden externen Bewertenden im Vergleich zu der Gruppe von Lehrkräften ist an vielen Stellen sehr hoch. Eine Begrenzung ist allerdings, dass externe Bewertende möglicherweise anders bewerten, wenn sie in einer Studie teilnehmen und eine Zweitkorrektur zu Bewertungen anderer Lehrkräfte unternehmen sollen. Wie in anderen Studien angesprochen (vgl. ibid.), könnte es sein, dass die schwedischen externen Bewertenden bei der Zweitkorrektur – bewusst oder unbewusst – strenger beurteilen. Es könnte andererseits durchaus auch der Fall sein, dass gerade diese beiden Bewertenden strengere Profile aufweisen.

Auch die Gruppe der Lehrkräfte könnte von der Teilnahme in der Studie beeinflusst sein. Praktizierende Lehrkräfte kennen die Kenntnisse ihrer eigenen Schülerinnen und Schüler und wissen außerdem, was sie im Unterricht behandelt haben. Dies kann in die Bewertungen der eignen Lernenden einfließen (vgl. Håkansson Ramberg 2021) und sie bewerten möglicherweise auch aus dem Grund milder, da sie wissen, dass die Lernenden sich für diese Prüfung

nicht vorbereitet haben.[183] Bei der Bewertung der eigenen Lernenden fließen womöglich auch frühere Leistungen im Laufe des Kurses in die Bewertungen ein. Ein Zeichen dafür sind Kommentare, die sich auf frühere Leistungen im Kurs beziehen (vgl. Kap. 6.5 unter der Kategorie *Sonstiges*). Ein weiterer Grund für milde Bewertungen könnte sein, was auch in anderen Studien zum Vorschein kommt, dass nämlich unterrichtende Lehrkräfte manchmal ihren Lernenden eine ausreichende Note (Note E) geben, da sie mit ihrer aus der Grundschule gewählten Sprache weitergemacht haben und sie sozusagen durch eine ausreichende Note belohnt werden. Dass die Lernenden in der vorliegenden Untersuchung von ihrer jeweiligen Lehrkraft belohnt werden sollten, lässt sich jedoch nicht behaupten.

Als weiterer Grund für die etwas mildere Bewertung durch die Gruppe der Lehrkräfte könnte aber auch das Format der Schülerleistungen benannt werden. Die Lehrkräfte haben handgeschriebene Texte bewertet, während die externen Bewertenden computergeschriebene Textproduktionen zur Beurteilung erhalten haben. Die Instruktionen der Lehrkräfte sagten, dass sie ihre Bewertungen so durchführen sollten, wie sie das immer tun, und daher haben viele der Lehrkräfte auch Korrekturen im Text oder am Textrand geschrieben. Es könnte aber angenommen werden, dass computergeschriebene Texte niedriger bewertet werden im Vergleich zu handgeschriebenen. Ein möglicher Grund dafür ist, dass Schreibfehler nicht so sehr auffallen, wenn sie per Hand geschrieben werden. Bisherige Studien haben jedoch keine so großen Differenzen wie im vorliegenden Fall gezeigt und dazu eine höhere Notenabweichung bei Leistungen im niedrigeren Bereich gefunden (vgl. Powers et al. 1994), was dafür spricht ist, dass dies nicht als einzige Erklärung angenommen werden kann. Vor diesem Hintergrund wäre natürlich eine weitere Untersuchung unter denselben Bedingungen aufschlussreich, in der alle Schülerinnen und Schüler am Computer schreiben. Dies war leider im Frühling 2017 noch nicht an allen teilnehmenden Schulen möglich, sollte aber bald durchführbar sein, da die Schulen in Schweden aktuell für die kommenden digitalen Prüfungsformate auf nationaler Ebene umstellen.

Eine andere Einstufung durch die eigenen praktizierenden Lehrkräfte als durch die externen Bewertenden muss nicht automatisch bedeuten, dass diese Einstufung nicht stimmen würde. Die eigenen Lehrkräfte haben die

183 In lediglich vier der Schülergruppen in der vorliegenden Arbeit hatten die Schülerinnen und Schüler Erfahrung mit Sprachzertifikatsprüfungen in Deutsch (vgl. Kap. 5.2).

Möglichkeit, ihre Benotung am Ende des Schuljahres auf Basis breiterer Unterlagen der Lernenden zu begründen und erkennen, wenn ihre Lernenden Schreibfehler machen. Diese Möglichkeit, die Bewertungen auch auf die bisherigen Leistungen zu stützen, sollte aber nicht in die Bewertungen einzelner Leistungen einfließen. Spuren einer solchen Sichtweise kommen jedoch in den Bewerterurteilen nur sehr selten vor. Inwiefern die höhere Benotung der Texte durch die Deutschlehrkräfte in dieser Studie aufgrund bisheriger Leistungen im Laufe des Kurses erfolgte oder ob die Lehrkräfte von einer positiven Entwicklung einzelner Lernender über das Schuljahr beeinflusst sind, lässt sich durch die vorliegende Studie nicht beantworten. In einer Auswertung des schwedischen Schulsystems durch den OECD wurde bereits 2011 die Bedeutung einer Erhöhung der Bewerterübereinstimmung bei von Lehrkräften bewerteten Tests betont. Hierbei wurden unterschiedliche Maßnahmen vorgeschlagen, wie eine Zweitkorrektur (z. B. durch eine andere Lehrkraft im Fach) und Fortbildungsaktivitäten (vgl. Nusche et al. 2011). Diese Untersuchung zeigt, dass auch heute noch ein Bedarf an einer erhöhten Bewerterübereinstimmung, besteht.

9.3 Inferenz der Extrapolation: Aspekte der kriterienbezogenen Validität

Die Inferenz zur Extrapolation (*extrapolation*) bezieht sich darauf, inwiefern ein Testergebnis als ein Indikator für die Sprachkompetenz der Lernenden in einer realen Weltsituation wahrgenommen werden kann. Hierbei sollten die Inferenzen zur Extrapolation hinsichtlich Angaben zum Niveau der Leistung (vgl. Kane 2002) untersucht werden. In diesem Fall sollen Nachweise des Erreichens oder Nicht-Erreichens eines bestimmten Sprachniveaus eingeholt werden, indem folgende Frage gestellt wird: In welchem Verhältnis stehen Bewertungen schriftlicher Sprachkompetenz nach schwedischen Bildungsstandards von Deutschlernenden am Gymnasium und GER-Bewertungen hinsichtlich eines B1-Niveaus des GER zueinander? Zur tentativen Zuordnung der fremdsprachlichen Schreibkompetenz in den untersuchten Schülerproduktionen zu einem GER-Niveau B1 wurden die Bewertungen zweier externer GER-Bewertender eingesetzt und mit den Bewertungen der schwedischen Bewertenden verglichen.

Die kriterienbezogene Validität bezieht sich auf das Verhältnis zwischen Testergebnis und einem externen Kriterium (*criterion*), von dem angenommen wird, dass es die gleiche Kompetenz ausdrückt. Weir (2005) unterscheidet hauptsächlich zwischen drei Typen von kriterienbezogener Validität: i) Vergleiche der Testergebnisse zweier Tests, die unterschiedlich konstruiert sind, aber

das gleiche Konstrukt prüfen, ii) Vergleiche der Testergebnisse zweier Versionen desselben Tests, oder iii) Vergleiche gegen externe anerkannte Rahmenmodelle, wie den GER. In dieser Studie wird vor allem auf den letzteren Aspekt fokussiert, indem die Bewertungen der schwedischen Bewertenden mit Bewertungen im Hinblick auf ein Referenzniveau eines externen Rahmenwerks (GER) verglichen werden. Wenn Aspekte der kriterienbezogenen Validität untersucht werden sollen, ist von Bedeutung, dass das externe Kriterium ein valides Maß für das zu messende Konstrukt ist (vgl. Weir 2005). Die Fremdsprachenstufen im schwedischen System stehen im Verhältnis zu den Referenzniveaus des GER und daher ist anzunehmen, dass die beiden Systeme ähnliche Konstrukte wahrnehmen und, durch die Orientierung der schwedischen Fremdsprachenstufen an den GER-Niveaus, auch ähnliche Kompetenzen anstreben. An diesem Punkt wird zudem angenommen, dass dieser hier verwendete Test des schriftlichen Ausdrucks genutzt werden kann, um das Erreichen oder Nicht-Erreichen eines sprachlichen B1-Niveaus schriftlicher Leistungen in einem schwedischen Schulkontext einschätzen zu können. Die Resultate dieser Analysen werden hierbei in erster Linie mit Studien aus einem schwedischen Schulkontext verglichen und diskutiert.

Sprachkompetenz bei Tyska 5 auf einem B1-Niveau?

Die Ergebnisse der vorliegenden Studie zeigen, dass Textproduktionen in Deutsch am Ende von *Tyska 5* auf einem B1-Niveau eingestuft werden können und dies ist insgesamt in der Studie bei der großen Mehrheit der untersuchten Lernproduktionen der Fall. Aus diesen Befunden ist zu entnehmen, dass Textproduktionen mit steigenden Fremdsprachenstufen in immer höherem Grad auf einem B1-Niveau einstuft werden. Die Tatsache, dass die GER-Bewertenden bis auf eine Ausnahme bei der Einstufung eines B1-Niveaus übereinstimmig sind, stärkt zudem die Reliabilität dieser Ergebnisse. Die Ergebnisse dieser Studie können des Weiteren einen Hinweis darauf geben, inwiefern auch Leistungen von schwedischen Schülerinnen und Schülern auf den Fremdsprachenstufen *Tyska 3* und *Tyska 4* das angestrebte GER-Niveau B1 erreichen. Diese Resultate indizieren zugleich auch, bis zu welchem Grad die nationalen Bildungsstandards Schwedens sich an dem GER orientieren, und erlauben somit auch vorläufige Aussagen auf einer Systemebene.

Die Befunde zeigen, dass die Textproduktionen der teilnehmenden schwedischen Schülerinnen und Schüler, die gemäß den schwedischen Bewertenden die Anforderungen für den Kurs *Tyska 5* erfüllt haben, überwiegend auf mindestens ein GER-Niveau B1 eingestuft wurden. Durch die vorliegende Studie

konnte demnach gezeigt werden, dass die Interpretation der Testergebnisse darauf hindeutet, dass schwedische Schülerinnen und Schüler die Anforderungen für das angestrebte GER-Niveau B1 im Schreiben erfüllen, auch wenn dabei auf Grund der relativ schmalen Datenbasis keine sog. „*strong claims*" gezogen werden können. In der vorliegenden Studie kommt es aber nur einmal vor, dass eine Textproduktion auf *Tyska 5,* die mindestens eine Note E hat, auch unter einem B1-Niveau eingeschätzt wurde. Wiederum gibt es auch Texte mit der Note F, die von den GER-Bewertenden auf einem Niveau B1 eingestuft wurden (vgl. Tab. 38). Dies indiziert, dass diese Schülertexte als grenzwertige Leistungen zu betrachten sind. Mögliche Gründe für diese unterschiedliche Bewertung bei grenzwertigen Leistungen können gemäß der qualitativen Analyse der Bewerterkommentare darin liegen, dass die Bewertenden zum Teil unterschiedliche Aspekte berücksichtigen und dass zudem schwedische Bewertende, zumindest in diesem Vergleich zweier grenzwertiger Schülerleistungen, eine gewisse Tendenz haben, sprachliche Mängel und die inhaltliche Erfüllung bei der Bewertung strenger als die GER-Bewertenden einzuschätzen. Die Tatsache, dass Schülerleistungen, die nahe an der Bestehensgrenze liegen, von erfahrenen Bewertenden bei der Benotung ein Ergebnis jenseits dieser Grenze erhalten können, zeigte sich aber auch in den Studien des nationalen Prüfungsmaterials von Erickson (2019).

Das schwedische System ist als auf Basisstandards basierend konzipiert, d. h. das Ziel ist, dass möglichst alle Leistungen der Lernenden das angestrebte Niveau erreichen (vgl. Kap. 3.1). Dies bedeutet, dass eine grenzwertige Leistung mit einer am niedrigsten bewerteten Note E mit dem absoluten Mindestniveau für ein erfülltes B1.2 zu vergleichen sein sollte. Daher kann die Variation bei der Einstufung von Textproduktionen niedrigerer Niveaus eher als eine logische Konsequenz verstanden werden (vgl. Erickson 2019). Variabilität unter Bewertenden ist zudem weder neu noch erstaunlich. Da aber schwedische Schülerinnen und Schüler am Gymnasium *Tyska 5* als Wahlfach belegen, liegt die Vermutung nahe, dass die Lernenden auf dieser Stufe mehrheitlich motiviert und lernbereit sind. Erstaunlicher ist es in diesem Zusammenhang, dass einige Schülerleistungen auf *Tyska 5* die erforderlichen Kompetenzen für die Stufe nicht aufwiesen. Hierbei ist aber zu beachten, dass nur eine sehr begrenzte empirische Stichprobe zur Verfügung steht und dass nur wenige der für die Studie erhobenen Leistungen eine nicht ausreichende Note F erhalten haben (6 von insgesamt 56 erhobenen Textproduktionen aus *Tyska 5*, vgl. hierzu auch Kap. 5.2).

Allerdings stehen die Resultate der vorliegenden Arbeit im Kontrast zu den Ergebnissen der ESLC-Studie (European Commission 2012b), der Studie von Aronsson (2020) sowie des TAL-Projektes (vgl. Granfeldt et al. 2019b). Diese

Studien haben jedoch allesamt die Fremdsprachenkenntnisse von Lernenden auf einem niedrigeren Niveau, dem GER-Niveau A2, untersucht. Gemäß ihren Untersuchungen weisen die schwedischen Schülerinnen und Schüler am Ende der schwedischen Grundschule generell etwas mangelnde Fremdsprachenkenntnisse auf, die zum Teil unter dem von Skolverket angestrebten Erwartungsniveau des GER liegen. Weder die ESCL-Studie noch die Studie von Aronsson haben Schülerleistungen in Deutsch untersucht. Nur das TAL-Projekt hat sich mit Fremdsprachenkenntnissen in Deutsch beschäftigt, in diesem Fall mit der mündlichen Kompetenz. Auch wenn die Schülerleistungen hinsichtlich der mündlichen Kompetenz in der Studie generell nicht das zu erwartende Sprachniveau erreicht haben, ist zu bemerken, dass die mündliche Kompetenz der Lernenden in Deutsch im Vergleich zu der in Französisch und Spanisch zumindest auf einem etwas höheren Niveau eingestuft wurde.

Keine der bisherigen Untersuchungen hat also gerade die schriftliche Kompetenz in Kombination mit Deutsch fokussiert. Die hier erwähnten Untersuchungen beziehen sich außerdem, wie erwähnt, hauptsächlich auf Schülerleistungen am Ende der Grundschule, d. h. zusammenfassend: die Lernenden befinden sich auf einem niedrigeren Niveau des Schulsystems, und die Studien haben vorwiegend Spanisch untersucht (nur das TAL-Projekt hat alle drei Schulsprachen fokussiert). Zum Teil standen außerdem andere Teile der Sprachkompetenz im Zentrum (die ESLC-Studie und die Studie von Aronsson haben zwar die schriftliche Kompetenz in ihren Untersuchungen behandelt, aber nur im Hinblick auf Fremdsprachenkenntnisse in Spanisch).

Die Ergebnisse dieser Untersuchung weisen allerdings Übereinstimmungen mit vorherigen empirischen Studien vom schwedischen Gymnasium auf, die ebenfalls Indikatoren für eine Relation zwischen einem angestrebten GER-Niveau und den entsprechenden schwedischen Fremdsprachenstufen gefunden haben (vgl. Tyllered 2002, Borger 2018). Die Tatsache, dass schwedische Schülerinnen und Schüler bei *Tyska 5* am Gymnasium das angestrebte Sprachniveau B1 im Fach Deutsch als Fremdsprache zu erreichen scheinen, ist im Einklang mit den Ergebnissen der Studien in Englisch von Tyllered (2002) und Borger (2018), die eine gute Übereinstimmung zwischen Prüfungsmaterial bzw. Leistungen in Englisch und angestrebten GER-Niveaus gefunden haben. Diese Studien nehmen aber im Unterschied zur vorliegenden Arbeit hauptsächlich das zurzeit aktuelle schwedische nationale Prüfungsmaterial für das Fach Englisch als Ausgangspunkt.

Die Befunde der beiden genannten Studien in Kombination mit dem Resultat der vorliegenden Untersuchung könnten somit ein Hinweis darauf sein, dass Schülerleistungen am Gymnasium in höherem Ausmaß als Leistungen in der

Grundschule die Kompetenzanforderungen in der jeweiligen Fremdsprache erfüllen. Eine mögliche Erklärung dafür könnte sein, dass die Lernenden am Gymnasium älter sind und damit in höherem Grad Verantwortung für ihre Studien übernehmen können. Des Weiteren scheint zu gelten, dass hauptsächlich motivierte Schülerinnen und Schüler am Gymnasium ihre Fremdsprache weiterlernen (vgl. Cardelús 2015). Sprachlernende am Gymnasium belegen zudem häufiger eine theoretische Ausrichtung, was mit sich bringen könnte, dass die Lernenden in größerem Ausmaß bereit sind, Zeit und Kraft in ihre Sprachstudien zu investieren. Dazu ist das Erhalten von Meritpunkten am Gymnasium in höherem Grad präsent als in der Grundschule. Als möglicher Grund für die mangelnden Kenntnisse von Spanisch in der ESLC-Studie wurde angeführt, dass die großen Klassengrößen in Spanisch eine Rolle spielen können und dass viele Lehrkräfte im Fach nicht die entsprechende Ausbildung haben (vgl. Riis & Francia 2013). In Schweden haben derzeit die Lehrkräfte am Gymnasium in höherem Ausmaß die Lehrerberechtigung als die in der Grundschule, was zudem für Lehrkräfte in Deutsch im Vergleich zu Lehrkräften in Spanisch in höherem Grad zutrifft (Skolverket 2019b). Die Deutschlehrkräfte der vorliegenden Studie waren alle ausgebildete Gymnasiallehrkräfte, was in diesem Zusammenhang eine Bedeutung haben kann.

Verhältnis schwedischer Bewertungen zu den GER-Bewertungen

Die vorliegende Arbeit beschäftigt sich zudem mit der Frage, inwiefern Textproduktionen von Lernenden auch auf den niedrigeren Stufen *Tyska 3* und *Tyska 4* bereits ein erfülltes B1-Niveau erreichen können, eine bislang vernachlässigte Frage. Hierbei ist festzustellen, dass Schülerleistungen, die auf *Tyska* 3 oder *Tyska 4* die höheren Noten erhalten haben, auch nach den GER-Bewerterurteilen das B1-Niveau erreichen. Gemäß den GER-Bewertungen liegen die entsprechenden Textproduktionen auf *Tyska 3* somit mindestens eine GER-Stufe über dem intendierten Mindestniveau dieser Stufe, dem A2-Niveau. Dieses Ergebnis bedeutet jedoch nicht, dass im Umkehrschluss alle Textproduktionen mit der höchsten Note A am Ende des Kurses auf *Tyska 3* und *Tyska 4* automatisch ein erfülltes B1.2-Niveau im Schreiben erreichen würden. Diese Frage lässt sich aufgrund der schmalen empirischen Datenlage nicht mit Sicherheit beantworten.

Dass ein relativ großer Anteil der untersuchten Leistungen auf *Tyska 3* und *Tyska 4* das B1-Niveau hinsichtlich der schriftlichen Kompetenz erreicht (7/20 bzw. 11/20, vgl. Tab. 36), ist aber nicht im Einklang mit den Ergebnissen der ESLC-Studie. Nur ein sehr geringer Anteil der schwedischen Lernenden des

Spanischen hat dort am Ende der Grundschule bei der schriftlichen Kompetenz ein höheres Niveau erreicht (vgl. European Commission 2012b). Da dieser Studie eine bewusste Textauswahl zugrunde liegt, sollte dies jedoch mit Vorsicht interpretiert werden. Die Tatsache, dass mehrere Schülerleistungen hinsichtlich der schriftlichen Kompetenz in Deutsch Anforderungen über dem zu erwartenden GER-Niveau erfüllen, könnte verschiedene Gründe haben. Wie bereits oben erwähnt, können Faktoren wie das Alter der Lernenden, die typologische Verwandtschaft des Deutschen mit den Schwedischen und die vorwiegend theoretischen Ausrichtungen am Gymnasium eine Rolle spielen, diesen Vermutungen müsste jedoch durch weitere Studien nachgegangen werden.

Die Studie nimmt auch Bezug auf weitere Aspekte der kriterienbezogenen Validität, indem die Testergebnisse der schwedischen Bewertungen in einem ersten Schritt mit einer zeitgleich erhobenen externen Variable (*criterion*), in diesem Fall den Bewertungen der GER-Bewertenden, verglichen wurden. Da die Berechnungen zeigen, dass die Bewertungen der schwedischen Bewertenden und der GER-Bewertenden stark korrelieren, ist zu vermuten, dass die jeweiligen Bewertungen auf ähnlichen Konstrukten basieren. In einem zweiten Schritt wurden die Aspektbewertungen der jeweiligen GER-Urteile mit den schwedischen Bewertungen korreliert. Auch wenn sich die Korrelationskoeffizienten der jeweiligen Bewertungsdimensionen untereinander kaum unterscheiden, scheinen die Aspektbewertungen zu *Strukturen* und zum *Wortschatz* im Vergleich zu inhaltlichen oder textstrukturellen Aspekten stärker mit den GER-Bewertungen zu korrelieren. Dies könnte auf einen stärkeren Fokus auf Aspekte in den Bereichen *formale Strukturen* und *Wortschatz* in den Bewertungen der schwedischen Bewertenden hindeuten (vgl. hierzu auch Kap. 6).

Die Ergebnisse der vorliegenden Studie im Hinblick auf die kriterienbezogene Validität sollten jedoch durch weitere empirische Analysen der Fremdsprachenkenntnisse von schwedischen Schülerinnen und Schülern am Gymnasium ergänzt werden, gewiss auch im weiteren Sinne durch Untersuchungen anderer Kompetenzen, wie rezeptiver Kompetenzen oder mündlicher Produktion und Interaktion, und weiterer Fremdsprachen. Man muss sich außerdem dessen bewusst sein, dass der verwendete methodische Ansatz zur Validierung schriftlicher Sprachkompetenzen in Deutsch auf eine relativ schmale empirische Materialbasis bezogen ist und daher eher tentative Schlüsse über die Beziehung zwischen dem intendierten Niveau des GER und dem eingeschätzten Niveau für die Schülerleistungen gezogen werden können. Darüber hinaus könnten zusätzliche Methoden zur Bestimmung der Sprachkompetenz von Lernenden verwendet werden, wie z. B. Vergleiche der Testergebnisse zweier Versionen desselben Tests, um auch andere Typen der kriterienbezogenen Validität in Betracht zu ziehen.

10. Schlussbemerkungen

Die vorliegende Arbeit bietet u. a. eine Zusammenstellung empirisch gegründeter Befunde zu verschiedenen Aspekten der Validität bei der Bewertung schriftlicher Sprachkompetenz. Dabei werden relevante Aspekte aufgegriffen und diskutiert, die für unterschiedliche Schritte der Bewertung in einem schwedischen Schulkontext von besonderer Bedeutung sind: die Konzeptualisierung der Bewertenden für das zu messende Konstrukt (*Konstruktvalidität: kognitive Validität und Kontextvalidität*), die Bewerterübereinstimmung der schwedischen Bewertenden (*Validität der Ergebnisermittlung*) und der Bezug schwedischer Schülerleistungen zu einem externen Referenzniveau (*kriterienbezogene Validität*). In diesem Kapitel werden die zentralen Ergebnisse der Arbeit zusammengefasst sowie abschließende Schlussfolgerungen für die Forschung und die Unterrichtspraxis gezogen. Die Beantwortung der eingangs gestellten Fragen bildet den Ausgangspunkt für ein abschließendes Fazit und es wird dabei auch auf die methodischen bzw. inhaltlichen Grenzen der vorliegenden Studie eingegangen (Kap. 10.1). Es folgen ein erster Ausblick auf mögliche weitere Forschungsperspektiven sowie ein zweiter Ausblick mit Überlegungen zu didaktischen Implikationen und der Relevanz dieser Befunde für die Bewertung von Textproduktionen in einem schwedischen Schulkontext (Kap. 10.2).

10.1 Fazit und Grenzen der Studie

Das übergeordnete Ziel der vorliegenden Studie war es, verschiedene Validitätsaspekte bei der Bewertung schriftlicher Schülerleistungen in der zweiten Fremdsprache am Gymnasium zu untersuchen. Die Bewertung sowie die Verwendung und Interpretation von Testergebnissen in einer zweiten Fremdsprache ist in einem schwedischen Schulkontext ein weitgehend vernachlässigtes Thema. Bewertungsdiskussionen im schwedischen Kontext haben bisher häufig die Bewertung schriftlicher Produktion fokussiert, wobei u. a. mangelnde Reliabilität Aufmerksamkeit erregt hat (vgl. Gustafsson et al. 2014). Aus diesem Grund war es ein wichtiges Desiderat, Bewertungen im Bereich Textproduktion sowohl von den eigenen Lehrkräften als auch durch externe Bewertende im Hinblick auf verschiedene Aspekte der Validität zu analysieren. Die Besonderheiten der vorliegenden Studie liegen im empirischen Material – basierend auf einem bereits an das B1-Niveau kalibrierten Test – und darin, dass sowohl schwedische Bewertende als auch externe GER-Bewertende für

die Untersuchung gewonnen werden konnten. Eine eigens für diesen Zweck erhobene empirische Datenbasis liegt der vorliegenden Arbeit zugrunde. In Kombination mit einer Orientierung an Mixed-Methods-Ansätzen, die sowohl qualitative als auch quantitative Analysen von Aspekten der Validität ermöglichen, besteht somit die Hoffnung, dass die Studie einen gewissen Beitrag im Hinblick auf die hier fokussierten Fragestellungen leisten kann.

Die Ergebnisse zur ersten Frage, welche sich auf die Konzeptualisierung der Bewertenden konzentriert, wurden in Kapitel 6 dargestellt und sollen Hinweise darauf geben, welche Aspekte in den Bewerterurteilen besonders relevant für die Beurteilung sind. Hierbei zeigen die Analysen ein breites Spektrum, das hauptsächlich Aspekte der linguistischen und der pragmatischen Kompetenzen, sowie in einem gewissen Ausmaß auch Aspekte der soziolinguistischen Kompetenz beinhaltet. In dieser Hinsicht konnten zudem Unterschiede zwischen schwedischen Bewertenden und GER-Bewertenden, die schwedische Bildungsstandards bzw. GER-Skalen bei der Bewertung verwendet haben, wahrgenommen werden. Diese Unterschiede können häufig auf die jeweiligen Bewertungskriterien zurückgeführt werden. Zum Teil sind die Unterschiede zwischen den Bewertergruppen auf das analytische bzw. holistische Bewertungsverfahren zurückzuführen. Die GER-Bewertenden beachten überwiegend Aspekte, die im zur Prüfung bereitgestellten Bewertungsraster zu finden sind, auch wenn dies nicht ausnahmslos zutrifft (z. B. für Aspekte zur *soziokulturellen Angemessenheit* bzw. zur *Orthographie,* die ebenfalls im Raster aufgeführt sind, aber nicht in demselben Ausmaß vorkommen), und berücksichtigen dabei häufig mehr Aspekte pro Schülerleistung.

Die schwedischen Bewertenden beachteten im Vergleich zu den GER-Bewertenden ein etwas breiteres Spektrum im Hinblick auf die beachteten Aspekte auf. Die in den Bewerterurteilen vorkommenden Aspekte werden jedoch nicht von allen Bewertenden berücksichtigt und werden zum Teil auch unterschiedlich gewichtet. Es ist folglich auch eine gewisse Variabilität unter den schwedischen Bewertenden. Des Weiteren lassen sich innerhalb der Gruppen gewisse Unterschiede erkennen: sprachliche Korrekturen werden in höherem Grad in den Urteilen der Gruppe der schwedischen Deutschlehrkräfte kommentiert, während die externen Bewertenden in etwas höherem Grad die Erfüllung der Aufgabe beachten. Hierbei ist jedoch anzunehmen, dass eine beträchtliche Variation innerhalb der Gruppe der schwedischen Lehrkräfte zu finden ist.

Der höhere Anteil von sprachlichen Anmerkungen unter den unterrichteten Lehrkräften könnte sich dadurch erklären, dass es eine Unterrichtstradition in formreichen Sprachen wie Deutsch gibt und dass die Lehrkräfte das prüfen

wollen, was von ihnen im Unterricht behandelt worden ist. Korrekturen zur sprachlichen Korrektheit sind bei einer Bewertung schriftlicher Kompetenz zudem leicht überprüfbar und gelten als weniger zeitaufwändig im Vergleich zu Einschätzungen zur Umsetzung der inhaltlichen Anforderungen. In einem kriterienorientierten System, wie das heutige System in Schweden, wird im Vergleich zu einem normorientierten die Interpretation der Anforderungen in den Bildungsstandards in höherem Grad den Lehrkräften überlassen. Dies bedeutet, dass die Lehrkräfte an einzelnen Schulen die Bewertungskriterien unterschiedlich interpretieren können und somit bei der Einstufung von Leistungen verschiedene Bewertungsrahmen haben.

Zusammenfassend können diese Erkenntnisse zum Verständnis dafür beitragen, *was* bei einer Bewertung schriftlicher Kompetenz Berücksichtigung findet. Zu einem gewissen Grad entscheidend für die Bewertung schriftlicher Kompetenz scheint somit die Konzeptualisierung des Konstrukts der jeweiligen Bewertenden, basierend auf den jeweiligen Bewertungskriterien. Ferner scheint auch das Bewerterverhalten eine Rolle zu spielen, wobei zugängliche Bewertungsraster die Bewertung stark beeinflussen könnten. Die große Variationsbreite zwischen den schwedischen Bewertenden ist aber auffällig. Um eine Generalisierung dieser Ergebnisse vornehmen zu können, sollte die vorliegende Studie u. a. durch zusätzliche Aufgabenstellungen und andere Teilnehmende ergänzt werden.

Die Tatsache, dass die Gruppe der schwedischen Deutschlehrkräfte in der vorliegenden Untersuchung eine gewisse Variabilität im Hinblick darauf aufweist, *was* sie in Textproduktionen von Lernenden berücksichtigen und *wie* sie die jeweiligen Aspekte bei der Bewertung schriftlicher Kompetenz gewichten, könnte verschiedene Gründe haben. Eine gewisse Variabilität gehört aber auch zum Konstrukt. Da sich die Lehrpläne, die Bewertungskriterien und nicht zuletzt die Lehrerausbildung in Schweden in der Vergangenheit mehrmals verändert haben, ist es nicht ungewöhnlich, dass die Lehrkräfte in einem unterschiedlichen Verhältnis zu den Kriterien stehen. Sie haben zudem unterschiedliche Erfahrungen mit Notensystemen und Skalen sowie damit, wie diese in der Praxis umgesetzt werden sollen. Diese unterschiedlichen Erfahrungen der Lehrkräfte könnten dazu führen, dass sie auch verschiedene Auffassungen darüber, was eine gute Schülerleistung kennzeichnet, haben können. Eine wichtige Implikation dieser Studie ist daher der Bedarf einer gemeinsamen Sichtweise schwedischer Lehrkräfte in Bezug darauf, wie fremdsprachliche Schreibkompetenz von Schülerinnen und Schülern bewertet werden soll. Dieser Bedarf gilt höchst wahrscheinlich generell und nicht ausschließlich für die schriftliche Kompetenz im Fach Deutsch. Fortbildungsmaßnahmen für

schwedische Lehrkräfte im schulischen Kontext wären daher dringend angezeigt, damit alle Lehrkräfte ein breites Spektrum der schriftlichen Kompetenz bei der Bewertung jeder Schülerleistung berücksichtigen und nicht gelegentlich vorwiegend sprachliche Korrekturen durchführen.

Zur Beantwortung der zweiten Frage, die sich auf die Bewerterübereinstimmung der schwedischen Bewertenden konzentriert und deren Ergebnisse in Kapitel 7 dargelegt wurden, sind Ermittlungen zum Konsens und zur Konsistenz durchgeführt worden. Die Ergebnisse zeigen deutlich, dass die Bewerterübereinstimmung der Bewertenden im Hinblick auf die Konsistenzwerte, vor allem bezüglich der Rangkomponente, im höheren Bereich liegt. Dahingegen fallen die ermittelten Konsenswerte bezüglich der Bewerterübereinstimmung eher im niedrigeren Bereich aus, vor allem zwischen der Gruppe der Lehrkräfte und einem der beiden externen Bewertenden. Hierbei kann zudem festgestellt werden, dass die Gruppe der Lehrkräfte – die Abweichungen gelten allerdings nicht alle Textproduktionen – im Vergleich zu den externen schwedischen Bewertenden eine Tendenz zu Milde aufweist. Die externen Bewertenden dahingegen stimmen in höherem Grad miteinander überein und zeigen eine leichte Tendenz zur Strenge. Des Weiteren können Bewerterprofile unter den Bewertenden wahrgenommen werden (vgl. hierzu Eckes 2008), wie z. B. die Neigung zu einer Zentraltendenz bzw. zum Vermeiden von Extremwerten. In diesem Zusammenhang interessant ist zudem, dass bei Bewertungen im mittleren Bereich eine größere Variation aufweisen und dass sie häufiger bei der nicht ausreichenden Benotung, im schwedischen System die Note F, übereinstimmen. Dies ist manchmal gegen die eigenen Erwartungen von Lehrkräften (vgl. Håkansson Ramberg 2016; 2021) und kann in manchen Fällen durch eine fehlende Aufgabe erklärt werden.

Trotz des uneinheitlichen Bildes der Bewerterübereinstimmung im Hinblick auf den Konsens zwischen den schwedischen Bewertenden, was womöglich auch mit der Subjektivität bei der Bewertung freier schriftlicher Produktion zu tun hat (vgl. Bachman et al. 1995; Eckes 2011), ist ein ähnliches Bild in weiteren Studien aus einem schwedischen Kontext zu finden (vgl. Skolverket 2020b). Erklärt werden können die unterschiedlichen Befunde hinsichtlich der Konsens- bzw. Konsistenzwerte durch das schwedische System für schulische Bewertung. Einzelne Lehrkräfte haben im schwedischen System eine große Verantwortung dafür, wie sie den Fremdsprachenunterricht gestalten und was sie im Unterricht behandeln. Sie sind aber auch dafür verantwortlich, wie sie die sprachliche Kompetenz bewerten sowie wie sie die Bewertungskriterien interpretieren und verwenden. Es hat sich hierbei gezeigt, dass Lehrkräfte im schwedischen System die Leistungen der eigenen Schülerinnen und Schüler gut einschätzen

und in Relation zueinander setzen können (vgl. SOU 1942:11; Johansson 2013). Schwieriger scheint es, wenn Leistungen aus verschiedenen Klassen und Schulen miteinander verglichen werden sollen. So, wie das schwedische System im Hinblick auf Bewertung und Benotung heute aufgebaut ist, kennen die Lehrkräfte hauptsächlich die Kompetenzen der eigenen Schülerinnen und Schüler. Schwedische Lehrkräfte haben häufig weniger Erfahrung damit, Leistungen aus anderen Klassen oder Schulen zu bewerten. Insbesondere könnte dies für Leistungen in einer Fremdsprache der Fall sein, da viele dieser Lehrkräfte an ihren Schulen allein im Fach sind. Ein weiterer Grund könnte darin liegen, dass das nationale Prüfungsmaterial nur fakultativ ist und daher nicht in allen Schulen verwendet wird. Die Bewertung schriftlicher Kompetenz ist aber eine komplexe Aufgabe, die sowohl fachliche Kompetenz als auch kollegiale Unterstützung verlangt.

Dies bedeutet wiederum nicht, dass die Konsens- und Konsistenzwerte im Hinblick auf die Bewerterübereinstimmung der Bewertenden nicht verbessert werden könnten. Ähnlich wie in früheren Berichten und Studien (z. B. Eckes 2008; Skolverket 2009) konnte nachgewiesen werden, dass Lehrkräfte individuelle Vorlieben haben und manchmal unterschiedliche Interpretationen der Bewertungskriterien vornehmen. Diskussionen werden darüber geführt, wie man die Reliabilität bei der Bewertung stärken kann, z. B durch digitale Werkzeuge. Es besteht allerdings die Gefahr, dass Bewertende mehr Aufmerksamkeit auf relativ einfach zu bewertende Leistungsmerkmale richten (vgl. Erickson & Åberg-Bengtsson 2012), wie beispielsweise Aspekte der grammatischen oder orthografischen Beherrschung, und dadurch eher qualitative Aspekte, wie die Umsetzung der inhaltlichen Anforderungen oder das Spektrum sprachlicher Mittel, vernachlässigen. Fortbildungsangebote zur Förderung fachlicher Kompetenz hinsichtlich der Bewertung schriftlicher Leistungen sollten daher das breite Spektrum sowie qualitative Aspekte der kommunikativen Kompetenz beleuchten und zu fachlichen Bewertungsdiskussionen von Schülerleistungen ermutigen. Zugleich scheint von Gewicht zu sein, dass Lehrkräfte genügend zeitliche Ressourcen für fachliche Bewertungsdiskussionen mit Kolleginnen und Kollegen erhalten und dass organisatorische Voraussetzungen geschaffen werden müssen, die Auseinandersetzungen und Diskussionen über Schülerleistungen anderer Klassen und Schulen mit Fachkolleginnen und -kollegen ermöglichen.

Die dritte Frage befasst sich mit der Beziehung schwedischer Bewertungen der schriftlichen Kompetenz nach nationalen Bildungsstandards in einem schwedischen Schulkontext auf *Tyska 3*, *Tyska 4* und *Tyska 5* zu einem erfüllten B1-Niveau des GER. Hierbei kann festgestellt werden, dass eine Mehrzahl der

Schülerleistungen, entgegen empirischen Studien in der zweiten Fremdsprache aus der Grundschule (vgl. European Commission 2012b; Granfeldt et al. 2019b; Aronsson 2020), die Anforderungen eines angestrebten B1-Niveaus des GER auf *Tyska 5* erfüllt. Wie bereits in Kapitel 8 angesprochen, indiziert dies, bis zu welchem Grade schriftliche Schülerleistungen im schwedischen Schulsystem zu einem Referenzniveau B1 des GER zuzuordnen sind. Darüber hinaus konnte festgehalten werden, dass auch ein relativ großer Anteil der Schülerleistungen auf den niedrigeren Stufen *Tyska 3* bzw. *Tyska 4* in zunehmenden Grade den Anforderungen auf einem B1.2-Niveau hinsichtlich der schriftlichen Kompetenz genügt. Hier ergänzen die Befunde die bisher eher begrenzte Forschungslage. Diese Befunde könnten ein erster Hinweis darauf sein, dass schwedische Schülerinnen und Schüler im Fach Deutsch eine höhere Kompetenz besitzen als z. B. in den Fächern Französisch oder Spanisch, was mit der typologischen Verwandtschaft des Deutschen mit dem Schwedischen zu tun haben könnte. Die Resultate indizieren ferner, dass Lernende am Gymnasium in einem schwedischen Schulkontext in höherem Ausmaß die Anforderungen hinsichtlich der GER-Niveaus erfüllen (vgl. Tyllered 2002; Borger 2018) als in der Grundschule. Dennoch wären in dieser Hinsicht weitere komplementierende Untersuchungen vonnöten.

Des Weiteren zeigen Korrelationsberechnungen der jeweiligen Bewertungen, dass die Bewertungen nach schwedischen Bildungsstandards stark mit den GER-Bewertungen korrelieren, was darauf hindeutet, dass sie auf einem ähnlichen Konstrukt basieren. Dies spricht wiederum für eine Beziehung zwischen den schwedischen Bildungsstandards und einem Referenzniveau des GER. Da der Referenzrahmen mittlerweile ein international flächendeckend verwendetes Referenzsystem geworden ist, sollte eine Diskussion über die Nutzung des GER sowie seine Beziehung zu den Bildungsstandards in einem schwedischen Kontext geführt werden. Wenn das schwedische System den GER als Bezugspunkt anwenden soll, setzt dies nicht nur Kenntnisse der jeweiligen Referenzniveaus voraus, sondern auch, dass die Lehrkräfte über die Inhalte im Referenzrahmen im Hinblick auf das Lernen, Lehren und Beurteilen reflektieren können und Unterstützung darin erhalten, wie diese Inhalte in Beziehung zu den schwedischen Bildungsstandards gesetzt werden könnten. Deshalb sollten fachdidaktische Weiterbildungen über die Grundlagen des GER sowie der weiteren Ausgaben zum GER (vgl. Council of Europe 2020) und derer Beziehung zu schwedischen Standards für Lehrkräfte in einer Fremdsprache angeboten werden, damit ein gutes Verständnis für die bildungspolitischen Zusammenhänge zwischen den schwedischen Bildungsstandards und dem GER unter Lehrkräften und anderen Bildungsakteuren geschaffen werden kann.

Im Hinblick auf das übergeordnete Ziel der vorliegenden Studie, verschiedene Validitätsaspekte bei der Bewertung schriftlicher Schülerleistungen am Gymnasium hinsichtlich der zweiten Fremdsprache zu analysieren, kann festgehalten werden, dass Nachweise im Hinblick auf Aspekte der Validität in verschiedenen Schritten des Bewertungsprozesses gefunden werden können. Dennoch können auch gewisse Defizite in Bezug auf die Validität bei der Bewertung schriftlicher Kompetenz identifiziert werden. Diese zeigen sich hauptsächlich bei der Konstruktkonzeptualisierung der schwedischen Bewertenden sowie bezüglich der Validität der Ergebnisermittlung. Hierbei finden sich Anhaltspunkte, dass auch ein kontinuierliches Angebot von Weiterbildungsmöglichkeiten an Lehrkräfte hinsichtlich Bewertung und Benotung im schwedischen Schulkontext von Nutzen sein könnte. Insbesondere scheint dies für die Bewertung freier Produktion, d. h. angesichts der mündlichen bzw. schriftlichen Kompetenz, relevant zu sein.

Abschließend deutet diese Studie auf positive Resultate im Hinblick auf die kriterienbezogene Validität hin. Die Tatsache, dass schwedische Bewertungen mit mindestens einer E-Note auf *Tyska 5* auch in der Regel von GER-Bewertenden auf das angestrebte Niveau B1 eingestuft werden, ist vom Qualitätsstandpunkt betrachtet sehr gut. Es gibt zwar einige grenzwertige Textproduktionen, aber bei einer empirischen Validierung von Bewertungen produktiver Kompetenzen ist jedoch eine gewisse Überlappung zu erwarten und eine perfekte Übereinstimmung mit einem externen Kriterium ist selten vorzufinden. Dies wird auch von Messick aufgegriffen: „But validity, except in extreme cases, is not an all-or-none question. On the contrary, it is a question of the *degree* to which evidence and rationales support the adequacy and appropriateness of interpretations and uses of scores." (Messick 1989a: 10, *Hervorheb. im Original*). Messick (1989b) weist auch darauf hin, dass eine Validierung ein kumulativer Prozess ist, wonach fast jeder Hinweis, der als Nachweis der Validität angesehen werden kann, von Bedeutung ist. Dieser kumulative Ansatz hinsichtlich der Erfassung von Nachweisen zur Validität wird auch von Weir (2005) vertreten.

Die vorliegende Studie zeigt außerdem, dass eine Kombination qualitativer und quantitativer Methoden einen Beitrag zum verbesserten Verständnis für eine Bewertung fremdsprachlicher Schreibkompetenz leisten kann. Die Methoden können zudem für eine Validierung im Hinblick auf das Erreichen bzw. Nicht-Erreichen eines GER-Niveaus verwendet werden. Vorzugsweise mit empirischen Belegen kann bestimmt werden, in welchem Ausmaß die Fremdsprachenkenntnisse schwedischer Schülerinnen und Schüler im Fach *Tyska* die Anforderungen eines angestrebten GER-Niveaus erfüllen. Des Weiteren hat

die vorliegende Untersuchung wichtige Erkenntnisse zur Bewertung aus einer Bewerterperspektive geben können, inklusive relevanter Nachweise zu verschiedenen Validitätsaspekten in einem schwedischen Schulkontext.

Im Kapitel zu Forschungsdesign und Methodik wurde bereits auf Begrenzungen der Studie eingegangen (Kap. 5.4). Es geht hier vor allem um Grenzen im Hinblick auf die *Stichprobe* der teilnehmenden Schulen, Lehrkräfte und Probanden, die *Anzahl der Bewertenden*, bestimmte Charakteristika des zugrundeliegenden *Tests des schriftlichen Ausdrucks*, *kontextgebundene* Faktoren sowie die *Analysemethoden*. Die *Stichprobe* ist sowohl im Hinblick auf die Schulen als auch auf Lehrkräfte und Probanden im Datensatz relativ begrenzt, was bereits im Methodikkapitel aufgegriffen wurde. Dazu kann die Frage gestellt werden, welche *Anzahl* von Bewertenden notwendig ist, um hinreichend zuverlässige und vergleichbare Aussagen über eine Bewertung treffen zu können. Es können durch die kleine Stichprobengröße keine „*strong claims*" über den Fokus der jeweiligen Bewertenden, die Bewerterübereinstimmung sowie das Sprachniveau schwedischer Schülerinnen und Schüler am Gymnasium im Fach Deutsch vorgenommen werden und die vorliegende Arbeit sollte daher durch zusätzliche Studien ergänzt werden.

Zudem haben die teilnehmenden Schülerinnen und Schüler nur einen *Test des schriftlichen Ausdrucks* durchgeführt. Durch die Verwendung eines für das B1-Niveau kalibrierten Tests sollte jedoch eine höhere Reliabilität der Studie gewährleistet werden. Auch wenn es im Rahmen der vorliegenden Studie gar nicht möglich war, eine Reihe mit mehreren Tests durchzuführen, können aus nur einem einzelnen Test dennoch keine weitreichenden Interpretationen abgeleitet werden. Eine solche Interpretation würde bedeuten, dass der Zielbereich (*universe of generalization*) zu eng definiert wäre und damit nicht genug Rücksicht auf unterschiedliche Aufgabenformate, Testereignisse und Kontexte genommen würde (Kane 2013: 18). Die Fragestellungen wurden hier dementsprechend nicht im Hinblick auf verschiedene Testformate, Testaufgaben und unter verschiedenen Testerreignissen untersucht. Einschätzungen über die schriftliche Kompetenz von Lernenden in einer Fremdsprache werden aber auch von Testinstituten häufig auf der Grundlage von einem einzelnen Test der schriftlichen Kompetenz gemacht. Darüber hinaus sind Sprachlernende aus mehreren schwedischen Schulen beteiligt und die Texte sind jeweils von zwei unabhängigen GER-Prüfern evaluiert worden. Eine Auswahl war im Rahmen dieser Studie notwendig, und alle Schülerinnen und Schüler umfassend zu testen wäre zudem unrealistisch gewesen.

Ebenfalls als Begrenzung sind zuletzt *kontextgebundene* Faktoren zu betrachten. Hierbei stellen die teilweise unterschiedlichen Voraussetzungen

der jeweiligen Bewertenden und Bewertergruppen im Hinblick auf zeitliche Ressourcen und das Bewertungsverfahren eine Begrenzung der Studie dar. Des Weiteren haben die Deutschlehrkräfte die Originalversionen der Schülerleistungen bewertet, während die externen Bewertenden eine digitalisierte Version erhalten haben. Nicht zuletzt als Teil der in den vergangenen Jahren erfolgten Vorbereitung für die Digitalisierung der nationalen Prüfungen in Schweden (vgl. dann Skolverket 2021d), aber auch aufgrund des entstandenen Bedarfs an digitalen Lösungen während der COVID-19-Pandemie, hat heute fast jede Schülerin und jeder Schüler an schwedischen Gymnasialschulen Zugang zu einem eigenen Computer für den Unterricht. Dies war zur Zeit der Datenerhebung im Frühjahr 2017 leider noch nicht immer der Fall. Eventuelle Replikationsstudien sollten daher erwägen, eine Datenerhebung mit digitalen Texten vorzunehmen, um sämtlichen teilnehmenden Bewertenden die gleichen Unterlagen zur Bewertung bereitstellen zu können. Auch wenn die vorliegende Untersuchung von kontextuellen Faktoren begrenzt wurde, liegt eine besondere Stärke im authentischen Schülertextkorpus. Hinsichtlich der Bewertung der Schülerleistungen sind mehrere und unterschiedliche *Analysemethoden* verwendet worden, um eine höhere Anzahl von Nachweisen zu erhalten und ein möglichst vielfältiges Bild von den verschiedenen Schritten einer Bewertung zu erhalten.

10.2 Ausblick: Weitere Forschungsperspektiven und didaktische Implikationen

Die Ergebnisse zeigen, dass die gewählten Methoden wertvolle Einsichten zu Aspekten der Validität in unterschiedlichen Schritten einer Bewertung vermitteln können, und zwar auch bei der vergleichsweise kleinen Stichprobengröße und einer eher geringen Anzahl von externen Bewertenden. Eine Besonderheit der vorliegenden Arbeit ist zudem die Breite und die Authentizität des empirischen Schülerkorpus sowie die Verwendung zweier unterschiedlicher Standards – der nationalen Bildungsstandards in Schweden und des europäischen Referenzrahmens – hinsichtlich einer Bewertung schriftlicher Kompetenz. Die vorliegende Untersuchung wirft aber weitere Fragen für zukünftige Studien auf.

So wurde beispielsweise nicht untersucht, welche individuellen Auffassungen und Voraussetzungen – neben den untersuchten Rahmenbedingungen – einen Einfluss auf die Bewertung haben. In der heutigen Gesellschaft ist es wichtig, eine Fremdsprache in authentischen Situationen verwenden zu können, was durch den Fokus eines handlungsorientierten Fremdsprachenunterrichts Konsequenzen sowohl für den Unterricht als auch für die Bewertung

sprachlicher Kompetenz gehabt hat. Ein in diesem Zusammenhang interessanter Aspekt wären die Auffassungen der Lehrkräfte, sog. *teachers' beliefs*, im Hinblick auf das Lernen, Lehren und Beurteilen einer Fremdsprache in einem schwedischen Schulkontext. Wenn allen Schülerinnen und Schülern in Schweden eine zuverlässig hochwertige schulische Ausbildung angeboten werden soll, sollte es von Gewicht sein, dass Lehrkräfte ein gemeinsames Verständnis für die Bewerterskriterien haben, aber auch, dass Lehrkräfte einer Fremdsprache eine gemeinsame Basis dafür entwickeln, *was* zu bewerten ist, und dass dem Fremdsprachenunterricht und der Bewertung ähnliche Prinzipien zugrunde liegen. Eine Untersuchung zum Einfluss kontextueller Faktoren wäre daher vonnöten: wie sich z. B. Unterrichtskontext sowie bisherige Erfahrungen der Lehrkräfte im Bereich der Bewertung und im Umgang mit den Bildungsstandards auf die Validität und Reliabilität bei einer Bewertung auswirken und in welcher Beziehung diese zu den *teachers' beliefs* in einem schwedischen Schulkontext stehen.

Die vorliegende Studie hat gezeigt, dass schwedische Lehrkräfte nicht nur verschiedene Aspekte bei der Bewertung berücksichtigen oder die gleichen Aspekte unterschiedlich gewichten, sondern, dass sie auch unterschiedliche Bewertungsstrategien verwenden. Einige nutzen eher analytische Bewertungsmatrizen, während andere eine holistische Bewertung basierend auf Bewertungskriterien oder kommentierten Schülerbeispielen verfolgen. Dies sollte kein Problem sein, solange die Bewertenden nicht irrelevante Kriterien in ihre Entscheidungen miteinbeziehen oder ihre Bewertungen auf eine zu selektive Auswahl von Kriterien gründen. Vor diesem Hintergrund wäre es in einer zukünftigen Studie von höchster Relevanz, zu untersuchen, welche Bewerterpraktiken Lehrkräfte in welchen Kontexten verwenden, wie signifikant sie sich unterscheiden und inwieweit diese unterschiedlichen Praktiken einen Einfluss auf die Validität und die Reliabilität bei einer Bewertung zu haben scheinen.

Nachweise für eine empirische Anbindung der Fremdsprachenstufen des schwedischen Systems an die Referenzniveaus des GER sind aus mehreren Gründen relevant: Erstens wäre dies als eine Qualitätssicherung hinsichtlich des schwedischen Systems zu betrachten, und zweitens würde dies ermöglichen, dass Lehrkräfte und Lernende sich in höherem Ausmaß am Referenzrahmen orientieren können. Der Referenzrahmen kann nicht nur im Hinblick auf die Bewertung oder darauf, Fremdsprachenkenntnisse eines bestimmten Sprachniveaus nachweisen zu können, hilfreich sein, sondern kann auch für die Verwendung von Lernmaterialien, für Testentwicklung und für Rater-Trainings von Bedeutung sein. Zukünftige Studien sollten sich daher mit der Zuordnung der Fremdsprachenstufen des schwedischen Systems zum GER

auseinandersetzen, auch im Hinblick auf andere Teile der Sprachkompetenz wie die rezeptiven Fertigkeiten sowie die mündliche Interaktion und Produktion. Vor allem scheinen Studien hinsichtlich der Fremdsprachenkenntnisse von Lernenden am Gymnasium vorrangig vonnöten, da der Fokus bisheriger Studien zu diesem Thema hauptsächlich auf der Grundschule liegt.

Des Weiteren erscheinen zudem Studien zum Sprachfertigkeitsniveau in einer Fremdsprache notwendig, da bisherige Studien aus der Grundschule im schwedischen Kontext generell auf mangelnde Sprachkenntnisse bezüglich der Sprachkompetenz der Lernenden in der zweiten Fremdsprache hingedeutet haben (vgl. European Commission 2012b; Granfeldt et al. 2019b; Aronsson 2020). Hierbei sollte, wie im TAL-Projekt in der Grundschule bereits geschehen, vorzugsweise eine landesweite Bezugsstudie zum GER vorgenommen werden, die die drei Fremdsprachen Deutsch, Französisch und Spanisch am Gymnasium in den Blick nimmt. Zu beachten ist jedoch, dass der GER einen Referenzpunkt für Bildungssysteme darstellt, jedoch kein überstaatliches Dokument ist und daher keine bildungspolitische „Zwangsjacke" der Länder werden darf (vgl. Kap. 2.3.3). Kritische Stimmen zur Notwendigkeit von regionalen und nationalen Anpassungen bei der Implementierung (z. B. North 2007) sollten daher nicht überhört werden. Es könnte hierbei eine Gefahr bestehen, dass Inhalte unkritisch übernommen werden und es ist daher für die Herstellung eines glaubwürdigen Bezugs zum GER von größter Relevanz, dass nicht nur textuelle Validierungsstudien vorgenommen werden, sondern dass die Ergebnisse dieser Studien auch empirisch untersucht werden.

Darüber hinaus darf nicht vergessen werden, dass man das Testen und Bewerten von Fremdsprachenkenntnissen nicht in einem geschlossenen System vornehmen kann und dass kontextuelle Faktoren und Konzepte beachtet werden müssen. Stobart (2003) formuliert dazu: „assessment is never a neutral process – it always has consequences. The task is to make these as constructive as possible, particularly for those who are assessed." (S. 140). An diesem Punkt sollten nicht nur die Interpretation und Verwendung der Testergebnisse, sondern auch ihre eventuellen Konsequenzen, sog. *washback effects*, für den Unterricht untersucht werden. Die Auswirkung einer Bewertung ist ein wichtiger Aspekt der Konsequenzvalidität (vgl. Weir 2005). Was wird im Fremdsprachenunterricht behandelt, in welcher Beziehung stehen Unterricht und Bewertung und welchen Einfluss hat womöglich der europäische Referenzrahmen auf das Lernen, Lehren und Beurteilen in einem schwedischen Schulkontext? Es ist anzunehmen, dass die Verwendung von Tests, die sich explizit am GER orientieren und mit deren Ergebnissen Fremdsprachenkenntnisse auf einem bestimmten GER-Niveau nachgewiesen werden, eine positive

Rückwirkung auf den Unterricht, das Fremdsprachenlernen und den Status der zweiten Fremdsprache in Schweden haben könnte. Inwiefern dies der Fall ist, sollte jedoch durch weitere Studien ergänzend untersucht werden.

Abschließend können aus den Ergebnissen der Studie Konsequenzen für die Unterrichtspraxis in einem schwedischen Schulkontext abgeleitet werden. Generell kann eine Reihe von didaktischen Schlussfolgerungen für den schwedischen Schulkontext, hauptsächlich im Bereich Bewertung, aber auch im Hinblick auf den Schreibunterricht, gezogen werden. Das Gewicht einer gemeinsamen Konzeptualisierung für das zu messende Konstrukt unter schwedischen Bewertenden ist durch die vorliegende Arbeit deutlich geworden. In einem schwedischen Schulkontext scheint es, u. a. durch die große Verantwortung der Lehrkräfte für die Gestaltung des Unterrichts und für die Bewertung, häufig vorzukommen, dass Lehrkräfte bei einer Bewertung schriftlicher Kompetenz individuelle Gewichtungen vornehmen und eigene Vorlieben haben. Beispiele solcher Unterscheide zwischen den Lehrkräften sind u. a. die Bewertung der inhaltlichen Aufgabenerfüllung und Aspekte der linguistischen Kompetenz. Dies führt dazu, dass Schülerleistungen im Hinblick auf verschiedene Fokusse der Bewertenden zum Teil unterschiedlich bewertet werden, und hierbei lässt sich auch fragen, inwieweit sämtliche Lernenden dieselbe Chance bei der Bewertung haben.

Es besteht die Hoffnung, dass eine Konsequenz der Studie ist, dass zukünftig ein breiterer Ansatz zur schriftlichen Kompetenz verfolgt wird, wonach nicht nur die leicht zu erfassenden Aspekte berücksichtigt werden. Es sollte dabei wichtig sein, eine Balance zwischen unterschiedlichen Teilen der schriftlichen Kompetenz zu schaffen, d. h. zwischen linguistischen, soziolinguistischen und pragmatischen Kompetenzen. Darüber hinaus ist von Gewicht, dass nicht hauptsächlich Defizite, sondern auch Qualitäten in den Textproduktionen beachtet werden. Eine solche Sichtweise ermöglicht den Lernenden, die Breite ihrer Kompetenz zu zeigen (vgl. Erickson 2020a) und eröffnet Potenziale für eine vielseitige Bewertung. Dies ist insbesondere auch dann ratsam, wenn eine formative Bewertung der Sprachverwendung von Lernenden vorgenommen werden soll. In der Rückmeldung an die Schülerinnen und Schüler sollten Stärken und Lernbereiche, die verbessert werden können, sowie Möglichkeiten, wie die Lernziele erreicht werden können, identifiziert werden. Obwohl nicht alle Aspekte, aus denen die kommunikative Kompetenz eines Individuums besteht, vollständig getestet werden können, und obwohl die Forschung sich womöglich niemals abschließend darauf einigen können wird, welche Komponenten diese Kompetenz umfasst, ist doch jeder Schritt in Richtung einer erhöhten Validität und Reliabilität hinsichtlich der Bewertung von größter Bedeutung. Eine Bewertung soll nicht nur relevant bezüglich des zu messende Konstrukts,

sondern auch möglichst zuverlässig sein, bei wiederholter Bewertung sollte man also zu demgleichen oder einem ähnlichen Ergebnis kommen. Zu bemerken ist zuletzt auch, dass ethische Aspekte bei der Bewertung beachtet werden müssen, damit die Lernenden in ihrem Lernprozess unterstützt und respektiert werden (vgl. Erickson 2020a). Welche Aspekte der schriftlichen Kompetenz Bewertende in ihren Urteilen von Schülerleistungen hervorheben, hat folglich einen Einfluss auf die Interpretation und Verwendung der Testergebnisse und könnte somit auch Konsequenzen für das Fremdsprachenlernen haben.

Um die Bewerterkompetenz unter den Lehrkräften zu erhöhen und um die Bewerterübereinstimmung und damit auch die Gleichwertigkeit zu fördern, ist es sehr wichtig, dass die Lehrerausbildung relevante Elemente im Bereich Bewertung enthält. Für viele der Lehrkräfte in einem schwedischen Schulkontext scheint die Auseinandersetzung mit Bewertung und Benotung etwas, das sie erst in ihrer Berufstätigkeit gelernt haben (vgl. Håkansson Ramberg 2016). Daher wäre es auch dringend angezeigt, dass verstärkt Bildungsangebote im Bereich Testen und Bewerten in der Lehrerausbildung des schwedischen Systems gegeben werden. Zusätzliche Fortbildungs- und Diskussionsforen für bereits berufstätige Lehrkräfte sollten aber ebenfalls nicht vernachlässigt werden, dies gilt nicht zuletzt für Lehrkräfte einer zweiten Fremdsprache, da sie in ihren Fächern an der Schule häufig die einzige Lehrkraft sind. Möglichst sollten Diskussionsgruppen nicht nur aus Lehrkräften einzelner Schulen, sondern auch aus Lehrkräften unterschiedlicher Schulen und Schultypen bestehen. Beispiele für bestehende Weiterbildungen auf höheren Ebenen sind die Module der sog. *Språksprånget* (2018b), die darauf abzielen, Lehrkräfte weiterzubilden, kollegiale Diskussionen zu fördern und Lehrerinnen und Lehrer im beruflichen Alltag zu unterstützen. Fortbildungsansätze im Bereich Bewertung könnten in der Verlängerung dazu beitragen, ein gemeinsames Verständnis für das zu messende Konstrukt zu schaffen und damit verbundene Differenzen und Ungleichgewichte bei der Bewertung zwischen Schulen, Gemeinden und Landesteilen zu überbrücken. Es ist zudem von Relevanz, dass zeitliche Ressourcen für Fortbildungsveranstaltungen und Bewertungsdiskussionen gegeben werden.

Zu beachten ist, dass Urteilstendenzen der Bewertenden zu mangelnder Bewerterübereinstimmung führen können. Wenn, wie in der vorliegenden Untersuchung gezeigt, deutliche Unterschiede zwischen den Konsens- bzw. Konsistenzwerten vorliegen, könnte dies auf Milde-Strenge-Differenzen deuten. Dieses Bild wird auch von den vorgelegten Ergebnissen aus Multifacetten-Rasch-Analysen und Kreuztabellen unterstützt. Inwiefern diese Ergebnisse als ein Effekt der Teilnahme an der Forschungsstudie (vgl. Gustafsson & Erickson 2013) zu betrachten sind, bleibt jedoch unklar. Um in einem schwedischen

Schulkontext weitere Aussagen über die Neigung von Bewertenden zu Urteilstendenzen, wie z. B. Tendenzen zur Milde bzw. Strenge, treffen zu können, sollten daher ergänzende Studien durchgeführt werden. Ein im Schulkontext in Schweden aktuell diskutiertes Thema ist die Frage, inwieweit das Testergebnis eines Individuums in den landesweiten Leistungstests im Rahmen der nationalen Prüfungen (vgl. Kap. 2.2.4) mit der Endnote im Fach korreliert. Schulen mit ähnlichen Testergebnissen weichen bei der Einstufung der Endnote unterschiedlich stark ab, was auf Tendenzen zur Strenge bzw. Milde auch bei der Vergabe der Endnote zwischen unterschiedlichen Schulen hindeutet (vgl. Skolverket 2020b). Inwiefern diese Unterschiede auch hinsichtlich der Beziehung zwischen den fakultativen Prüfungen und der Endnote im Fach *Moderna språk* zu finden sind, bleibt unklar und ist zudem schwerer zu untersuchen, da diese Tests im gegenwärtigen System nicht obligatorisch sind.

Die vorliegende Studie zeigt, dass die eigene Lehrkraft und ein externer Bewertender bei der Bewertung derselben Schülerleistung häufig zu unterschiedlichen Testergebnissen kommen. Bisherige Studien zeigen, dass Bewertungen durch zwei Bewertende zu bevorzugen sind, um die Bewerterübereinstimmung zu erhöhen (vgl. Skolinspektionen 2018; Dalberg 2019). Dies kann entweder durch ein Verfahren mit externen Bewertenden oder durch sog. „*sambedömning*" (etwa ein paralleles Bewertungsverfahren) implementiert werden. *Sambedömning* setzt voraus, dass die Lehrkräfte die Leistungen diskutieren und diese nach den Anforderungen in den Bildungsstandards bewerten (vgl. Skolinspektionen 2018), eine Vorgehensweise, die mehrere Vorteile hat: Wenn Lehrkräfte die Bewertung von Schülerleistungen gemeinsam diskutieren, kann hoffentlich ein gemeinsames Verständnis für die Interpretation und Verwendung der Kriterien geschaffen werden. Lehrkräfte, die an solchen Bewertungsgesprächen teilnehmen, halten es für plausibel, dass der Grad an Bewerterübereinstimmung somit steigt (vgl. Connolly et al. 2012). Dies liegt wahrscheinlich daran, dass die Lehrkräfte durch *sambedömning* dazu neigen, ein gemeinsames Verständnis für das zu messende Konstrukt, d. h. *was* zu bewerten ist, zu entwickeln. Mit einer stärkeren Emphase auf Bewertungsdiskussionen ist zu vermuten, dass auch die Bewerterkompetenz unter den teilnehmenden Lehrkräften zunimmt. Darüber hinaus können durch *sambedömning* erfahrene Lehrkräfte weniger erfahrene Lehrkräfte bei der Bewertung unterstützen. Bisherige Studien haben zudem bereits gezeigt, dass Verhandlungen zwischen Bewertenden zu positiven Effekten führen könnten, sowohl im Hinblick auf den Beurteilungsprozess als auch auf die Gestaltung der Unterrichtspraxis (vgl. Trace et al. 2017). Diese Ergebnisse verweisen somit auf mehr Vorteile einer Fachdiskussion zwischen Lehrkräften als lediglich die adäquate schulische Leistungsbeurteilung.

Im Kontext eines Bewertungsverfahrens mit zwei Bewertenden wäre zudem von größter Relevanz, dass eine Organisation geschaffen würde, die eine breite Auswahl von Schülerleistungen berücksichtigt und die Bewertungsdiskussionen nicht nur zwischen Lehrkräften einzelner Schulen, sondern auch zwischen Lehrkräften aus verschiedenen Schulen und Schultypen ermöglicht. Wenn aber aus praktischen oder zeitlichen Gründen nicht alle Schülerleistungen durch zwei Bewertende beurteilt werden können, ist des Weiteren von Gewicht, dass nicht nur grenzwertige Schülerleistungen im unteren Bereich von den Lehrkräften diskutiert werden, sondern gerade solche Leistungen, bei denen die Lehrkräfte das Ergebnis als eindeutig einschätzen und die sich auch im mittleren oder höheren Bereich befinden.

Zusammenfassend lässt sich konstatieren, dass die vorliegende Untersuchung vielfältige Befunde im Hinblick auf die Bewertung schriftlicher Schülerleistungen in einem schwedischen Schulkontext und darauf liefert, wie diese zu einem bestimmten GER-Niveau zuzuordnen sind. Empirische Erkenntnisse über eine Anbindung der Fremdsprachenstufen des schwedischen Systems an die Referenzniveaus des GER haben didaktische Implikationen für den Fremdsprachenunterricht. Die Tatsache, dass auch Textproduktionen von Schülerinnen und Schülern auf niedrigeren Fremdsprachenstufen, in diesem Fall *Tyska 3* und *Tyska 4*, die Anforderungen eines B1-Niveaus im Hinblick auf die schriftliche Kompetenz erfüllen, weist darauf hin, dass die Fremdsprachenkenntnisse der Lernenden bereits auf niedrigeren Stufen weit über dem angestrebten Niveau liegen können, was auch im Unterricht berücksichtigt werden sollte. Hinsichtlich der Verwendung des Referenzrahmens in einem schwedischen Schulkontext scheint es zudem von großer Bedeutung zu sein, dass der Einfluss des GER auf das schwedische Schulsystem näher beleuchtet wird und dass klar wird, in welcher Beziehung der GER zu den schwedischen Bildungsstandards steht. Eine zukünftige empirisch validierte Anbindung des schwedischen Fremdsprachenstufensystems an den Referenzniveaus des GER setzt jedoch nicht nur voraus, dass Lehrkräfte mit den Referenzniveaus vertraut sind. Sie würde auch erfordern, dass Lehrkräfte dem Dokument nicht einfach unkritisch gegenüberstehen, sondern über den Inhalt des GER reflektieren können. Inwieweit schwedische Lehrkräfte sich an den Referenzniveaus des GER orientieren und in welchem Ausmaß sie den Referenzrahmen kennen, ist jedoch fraglich (vgl. Kap. 2.4). In der Lehrerausbildung und in weiteren Fortbildungsangeboten ist aus diesem Grund eine kritische Auseinandersetzung mit dem Inhalt des Referenzrahmens und seiner Beziehung zu den Bildungsstandards im Hinblick auf den Fremdsprachenunterricht und die Bewertung sprachlicher Kompetenz anzustreben.

Svensk sammanfattning

Inledning

Alltsedan digitaliseringen har flera stats-, nations- och andra gränser förlorat i betydelse. Detta leder till nya språkliga utmaningar och ett stort behov av språkkompetens, i såväl engelska som i andra främmande språk. Vikten av språklig kompetens i främmande språk lyfts idag även fram i olika riktlinjer och språkpolitiska dokument (t.ex. Skolverket 2018a; Council of Europe 2020). Dessutom betonas ofta nödvändigheten av att kunna kommunicera på minst två språk utöver modersmålet (jfr European Council 2002) och diskussioner förs hur man kan främja detta på olika sätt. Att kunna kommunicera på ett annat språk ses ofta som det främsta målet med språkinlärning och har haft till följd att språkundervisningen under de senaste årtiondena i allt högre grad kommit att inriktas mot kommunikativ kompetens. Under senare år har Europarådets *Gemensam europeisk referensram för språk: lärande, undervisning och bedömning* (2001), GERS, haft ett stort inflytande på språkundervisningen och blivit ett viktigt referensverktyg för hur språk lärs in, undervisas och bedöms. GERS ligger numera som referenssystem ofta till grund för bedömning i språk, främst inom länder i Europa, men även i andra delar av världen. Referensnivåerna i GERS används numera i allt högre grad när det gäller att definiera inlärares språkförmåga, vilket har lett till att allt fler språkinstitut, språkskolor, förlag och nationella utbildningssystem relaterar språkprov, språkkurser, kurslitteratur och styrdokument till dessa språknivåer. Inom Europa relaterar därmed bedömning av inlärares kompetens i främmande språk, såväl i utbildningskontexter som i yrkessammanhang, i allt högre grad till detta externa referenssystem.

Även inom svensk skola betonas ett kommunkativt synsätt i språkundervisningen. Till skillnad från engelska sker en stor och viktig del av språkinlärning och tillägnande av ett s.k. *modernt språk* inom en svensk kontext i skolan. Detta innebär att lärare i moderna språk spelar en mycket viktig roll för elevers lärande. Svenska lärare har också jämförelsevis en hög grad av autonomi när det gäller hur undervisningen ska utformas, vilket också innebär att de även i hög grad är ansvariga för bedömningen. De behöver alltså inte enbart besitta tillräckliga ämneskunskaper, utan behöver också ha en förtrogenhet och förståelse för kursmål, bedömningskriterier samt syfte, former och konsekvenser av en bedömning. För att bedömningar ska uppfylla sitt syfte behöver de vara giltiga (*valida*). Det handlar i språk ofta om att bedömningen så effektivt som möjligt

ska kunna fånga elevers språkliga förmåga. För att kunna undersöka validitet vid bedömning av elevtexter i ett modernt språk är det angeläget att närmare granska vilka delar av språkförmågan som lyfts fram vid en bedömning. Det kan dessutom vara viktigt att studera i vilken mån elevers språkkunskaper kan relateras till en bestämd språklig nivå och därmed även till språkanvändning i olika typer av verkliga situationer.

Bedömning och betygsättning har under senare år granskats och diskuterats inom en svensk utbildningskontext, såväl inom forskningen som från skolmyndigheternas sida (t.ex. Erickson 2009; Skolinspektionen 2010; Skar 2013, Borger 2018; Skolverket 2020b). Framför allt har dessa undersökningar handlat om nationella prov, bl.a. bedömarsamstämmighet i enskilda provdelar, skillnader mellan lärarbedömningar och externa bedömningar samt förhållandet mellan kursbetyg och elevers resultat vid nationella prov. Inom fältet för moderna språk har även studier berört huruvida elevprestationer inom det svenska utbildningssystemet uppnår språkliga referensnivåer enligt GERS (t.ex. European Commission 2012b; Granfeldt et al. 2019b; Aronsson 2020). I de svenska styrdokumenten uttrycks en tydlig koppling till referensnivåerna i GERS och samtliga språksteg i det svenska sammanhållna sjustegsystemet för språk (kurs 1–7) är relaterade till en bestämd referensnivå (A1–C2) enligt GERS (Skolverket 2011b). I den svenska kontexten relaterar exempelvis kurs 4 till en B1.1-nivå och kurs 5 till en B1.2-nivå enligt GERS. Även om de svenska ämnesplanerna i språk följaktligen är tydligt influerade av den europeiska referensramen, finns däremot relativt få empiriska studier som har undersökt detta samband (t.ex. Erickson & Pakula 2017), särskilt när det gäller de högre språkstegen i moderna språk på gymnasienivå.

Trots ett ökat intresse för bedömning och frågor om validitet och likvärdighet har förhållandevis få tidigare forskningsstudier gällt bedömning i skolämnet moderna språk, särskilt i tyska. Följaktligen finns ett stort behov av empiriska studier av lärarbedömningar av elevprestationer i tyska för att undersöka möjligheter och utmaningar vad beträffar validitet i en svensk skolkontext. Då bedömning av inlärares fria textproduktion visat sig kunna ge upphov till subjektivitet och olika tolkningar hos bedömare (t.ex. Skolinspektionen 2010; 2018) samt då en tidigare studie visat att svenska elevers prestationer i skrift (spanska) i en internationell jämförelse inte uppnår den förväntade språkliga referensnivån enligt GERS i moderna språk (European Commission 2012b), är det särskilt angeläget att undersöka bedömning av elevers skriftliga kompetens, d.v.s. skriftlig interaktion och produktion. Mot denna bakgrund är syftet med studien att belysa några validitetsaspekter när slutsatser dras av en bedömning av svenska elevers skriftliga kompetens på tyska i en svensk skolkontext.

Syfte och frågeställningar

Föreliggande studie tar därmed sin utgångspunkt i bedömning av elevprestationer inom en svensk skolkontext. Syftet med studien är att undersöka centrala validitetsaspekter vid en bedömning av elevers skriftliga språkkompetens i tyska i kurserna *Tyska 3*, *Tyska 4* och *Tyska 5* på gymnasienivå. Mer specifikt studeras a) bedömares fokus vid bedömning av elevers skriftliga kompetens, b) svenska bedömares samstämmighet, samt c) relationen mellan bedömningar av elevers skriftliga språkfärdighet på olika språksteg i den svenska stegmodellen och en bestämd språklig referensnivå enligt GERS. Mot denna bakgrund formulerades följande frågeställningar:

1. Vilka aspekter av inlärares skriftliga kompetens fäster bedömare särskilt avseende vid i sina bedömningar och hur skiljer sig dessa bedömningar åt mellan enskilda bedömare och bedömargrupper vad beträffar a) undervisande lärare, b) externa svenska bedömare samt c) GERS-bedömare?
2. Hur skiljer sig bedömningar åt beträffande bedömarsamstämmigheten mellan svenska bedömare?
3. Vilken relation har bedömningar av svenska gymnasieelevers skriftliga kompetens i kurserna *Tyska 3*, *Tyska 4* och *Tyska 5* i det svenska utbildningssystemet till bedömningar av skriftlig kompetens på en uppfylld B1-nivå enligt GERS?

Konceptuell ram

Definition, koncept och modeller för kommunikativ språkkompetens (jfr Hymes 1972; Bachman & Palmer 1996) ligger till grund för den handlingsorienterade språksyn som präglar både de svenska ämnesplanerna i moderna språk och GERS som utgör referenspunkt för det svenska systemet. För att undersöka validitetsaspekter av bedömning i en svensk kontext behöver begreppet *validitet* förklaras. Sedan mitten av 1900-talet har validitetskonceptet utvecklats och förändrats. I den traditionella indelningen delades validitet in i tre olika typer (jfr Messick 1989a): innehållsvaliditet (hur väl provinnehållet innehåller ett representativt urval av det provet avser att pröva), kriterierelaterad validitet (hur väl provresultatet kan relateras till ett externt kriterium, t.ex. resultatet från andra prov eller framtida kompetensnivåer) samt konstruktvaliditet (hur väl provet mäter de egenskaper eller den förmåga som det avser att mäta). I den idag dominerande definitionen utgår man ofta från ett enhetligt validitetskoncept där fokus istället ligger på provresultatets tolkning och användning. Enligt Messick (1989b) är validitet ett mångfacetterat begrepp där bedömningens två

funktioner, tolkning och användning, kan stödjas genom empiriska belägg och teoretisk underbyggnad. Bedömningar kan även medföra olika konsekvenser samt påverka uppfattningar och värderingar. Enligt Messick (jfr 1989a; 1995) finns två hot mot validiteten vid tolkning och användning av provresultatet, s.k. *construct-irrelevant variance* (provet innehåller dimensioner som inte tillhör det som provet avser att mäta) och *construct underrepresentation* (provet innehåller för få relevanta dimensioner av det som provet avser att mäta).

Under de senaste årtiondena har inom fältet för språkbedömning ett flertal teoretiska ramverk för validering presenterats. Messicks enhetliga validitetskoncept återfinns i flera teoretiska modeller för validering inom fältet för språkbedömning, t.ex. Kanes argumentbaserade ansatser (t.ex. Crooks m.fl. 1996; Kane 2006; 2013; Chapelle m.fl. 2008) och Weirs sociokognitiva modell (2005). I föreliggande studie används delar av Kanes validitetsmodell innehållande en kedja av inferenser (se kap. 3.2.2, figur 5), utvecklad att erbjuda ett stöd för vilka typer av belägg som behövs för att utvärdera tolkning och användning av provresultaten. Inferenskedjan består av följande delar: bedömning (*scoring*), generalisering (*generalization*), extrapolering (*extrapolation*) samt beslut (*decisions*). Studien använder även Weirs sociokognitiva modell (2005) som innehåller komponenter som bör utvärderas vid en validering (se kap. 3.2.2, figur 6). Enligt Weir kan validitet delas in i följande validitetsaspekter: kontextvaliditet (*context validity*), kognitiv validitet (*cognitive validity*), bedömningsvaliditet (*scoring validity*), kriterierelaterad validitet (*criterion-related validity*) samt konsekvensvaliditet (*consequential validity*). För föreliggande undersökning har resonemangen och redskapen i såväl Kanes inferenskedja som i validitetsaspekter hos Messick och Weir varit användbara vid analyserna av materialet och för vad som räknas som en valid bedömning.

Forskningsdesign

Forskningsdesignen orienterar sig mot en s.k. *mixed-methods*-ansats (t.ex. Kuckatz 2014a), vilken möjliggör användningen av såväl kvalitativa som kvantitativa metoder. Metoden innebär att inte enbart produkten av en bedömning undersöks utan även att bedömarnas förståelse av konstruktet vid bedömning av skriftlig förmåga analyseras, vilket även kan sägas ge en viss inblick i själva bedömningsprocessen.

Datainsamling

Materialet bygger på bedömningar av 60 för syftet särskilt insamlade elevtexter skrivna av svenska gymnasieelever i kurserna *Tyska 3*, *Tyska 4* och

Tyska 5 enligt den svenska språkstegmodellen (motsvarande ungefär GERS-nivåerna A2.2, B1.1 och B1.2). Tillvägagångssättet vid datainsamlingen kan bäst beskrivas som s.k. *purposive sampling*, där skolor och elevgrupper valts ut för att kunna fylla ett specifikt behov (jfr Robson & McCartan 2016). Skolornas rektorer kontaktades i ett första steg och varje skolas undervisande lärare i tyska i ett andra steg. I detta sammanhang finns dessutom inslag av möjlighetsurval; även om många var positiva till studien avböjde flera rektorer och lärare att medverka.

Provmaterialet för studien härrör från en vid provtillfället ännu inte offentliggjord modul av Goethe-institutets språkcertifikat för tyska på en B1-nivå enligt GERS (bilaga 9). Det faktum att provet innehöll flera olika uppgifter ökar reliabiliteten i bedömningen, vilket medför att effekten av varje enskild uppgift minskar och att det blir säkrare att generalisera från elevprestationen (jfr Weir 2005). Totalt 225 elevtexter, skrivna av elever från sammanlagt 24 grupper i tyska och 18 olika skolor, samlades in våren 2017 och fanns till förfogande för studien. Därefter gjordes ett urval av 60 elevtexter, 20 elevtexter från var och en av de tre kurserna *Tyska 3*, *Tyska 4* och *Tyska 5*. Det handlade här om ett representativt proportionellt stratifierat urval av texter med inslag av slumpmässighet i den mån att urvalet av texterna följde vissa på förhand klarlagda principer (se kap. 5.2, tabell 12).

För att svara mot studiens syfte bedömdes elevtexterna av i) de undervisande gymnasielärarna ii) två erfarna svenska bedömare samt iii) två externa, särskilt utbildade, GERS-bedömare. Den första gruppen bestod av undervisande lärare i tyska från både kommunala och fristående skolor i södra och mellersta Sverige. I den andra gruppen återfanns två externa svenska bedömare. Dessa båda bedömare hade dessutom på olika sätt under sitt yrkesliv samlat ytterligare erfarenhet av bedömning av elevtexter på tyska. Den tredje gruppen bestod av två certifierade GERS-bedömare med särskild erfarenhet och utbildning av att bedöma inlärares språkliga kompetens i tyska enligt GERS (jfr bilaga 10).

Bedömargrupperna använde sin egen bedömningsprocedur; de svenska bedömarna använde kunskapskraven i de svenska nationella ämnesplanerna i *Moderna språk* och därmed också den sex-gradiga skalan med betygsstegen F-A medan GERS-bedömarna genomförde bedömningen med hjälp av kriterier grundade i GERS. GERS-bedömarna använde sig av mer analytiskt inriktade bedömningskriterier för att avgöra om elevprestationen uppnådde en B1-nivå. Samtliga bedömare gav även en skriftlig motivering till sin bedömning av varje elevtext. Samtliga deltagare i studien informerades om undersökningens syfte och gav sitt samtycke till att medverka i studien enligt de forskningsetiska principer som Vetenskapsrådet (2002) ställt upp.

Analys

Kvantitativa och kvalitativa metoder har tillämpats för att analysera dels provresultat, dels skriftliga kommentarer till bedömningen och därmed kunna undersöka bedömningar av elevers skriftliga språkförmåga genomförda av de olika bedömargrupperna. De totalt 300 skriftliga bedömarkommentarerna analyserades huvudsakligen genom kvalitativ innehållsanalys. Ett tematiskt kodningsschema utvecklades genom både ett induktivt och deduktivt angreppssätt (jfr Kuckartz 2014b). Kodningsschema och kodning validerades i flera steg med hjälp av två oberoende medkodare, båda på universitetsnivå med erfarenhet av språk och bedömning. Totalt kodades upp till en tredjedel av materialet med en bedömarsamstämmighet för medkodarna och forskaren mellan 86 % och 94 %. Kodningsförfarandet erbjuder en kvalitetskontroll och stärker därmed reliabiliteten vad beträffar bildandet av kategorier i kodningsschemat och själva kodningen. Datorprogrammet *NVivo 12* användes för att strukturera och analysera materialet. För att ytterligare belysa resultaten beträffande den andra och tredje frågeställningen genomfördes jämförande analyser mellan bedömarkommentarer som bygger på den ovan beskrivna kvalitativa innehållsanalysen. Här valdes bedömningar till elevtexter med samma eller avvikande bedömning ut för att på så sätt kunna belysa skillnader och likheter i bedömnigarna.

Bedömarsamstämmigheten mellan de svenska bedömarna analyserades med hjälp av deskriptiv statistik och metoder för bedömarsamstämmighet. Att beakta är att varje statistisk metod har bestämda egenskaper vilket medför att beräkningarna ger olika typer av information. Av den anledningen användes olika beräkningar, såväl konsensus- som konsistensmetoder samt en Rasch-analys, för att nå en bredare bild av materialet. Även vid analysen av relationen till en extern referensnivå enligt GERS användes deskriptiv statistik och korrelationsberäkningar (Spearman’s Rho). För de kvantitativa analyserna användes datorprogrammen *SPSS* samt *MINIFAC*.

Resultat

De huvudsakliga resultaten av den empiriska studien fördelat efter de tre forskningsfrågorna presenteras nedan.

Analys av bedömares fokus vid bedömning av inlärares skriftliga kompetens

Analysen visar att svenska bedömare fäster avseende vid ett brett spektrum av aspekter i sin bedömning, där de mer lingvistiska aspekterna som *formella*

strukturer, *ordförråd* och en *övergripande språklig bedömning* sammantaget tycks vara något mer framträdande. Även aspekter som förmåga att *anpassa språket*, t.ex. till sociala konventioner, samt dimensioner som *fullgörande av uppgiften*, *begriplighet* och *helhetsintryck* kommer ofta till uttryck i kommentarerna. Vidare tycks en bedömning till stor del påverkas av bedömningsskalor och bedömningsförfarandet, d.v.s. om bedömningen är mer holistisk eller mer analytisk inriktad. Detta märks inte minst när det gäller GERS-bedömarna där kommentarerna i hög grad återspeglar kriterierna i bedömningsschemat. Därtill verkar GERS-bedömare i sina kommentarer ha en mer balanserad fördelning av de aspekter som lyfts fram som relevanta vid en bedömning, medan svenska bedömare ofta fäster avseende vid olika aspekter i sina bedömningar och i viss mån även viktar dessa olika.

Analys av bedömarsamstämmighet

Analyserna av samstämmigheten mellan de svenska bedömarna visar att konsistensvärdena är högre än konsensusvärdena. Detta tyder på att de svenska bedömarna i högre grad överensstämmer när det gäller rangordningen i bedömningen än ger en exakt överensstämmelse. Vidare visar analysen att de båda externa bedömarna i högre grad överensstämmer i sina bedömningar än jämförelser med de undervisande lärarna. Dessutom visar analysen att de svenska bedömarna är mer överens vid elevprestationer som erhåller lägre betyg jämfört med betygen i mitten eller högre betyg. Den kompletterande Rasch-analysen för de svenska bedömarna anger att de undervisande lärarna i jämförelse med de båda externa bedömarna generellt har en tendens till en något mildare bedömning. Avslutningsvis visar analysen att svenska bedömare oftare fäster avseende vid samma aspekter vid en bedömning som ger ett icke godkänt betyg F än vid högre betygssteg. Bedömare tycks därmed i högre grad göra liknande tolkningar av lägstakraven för en elevprestation. Däremot verkar bedömare vid högre betygssteg vikta aspekter olika, även om de alltså fäster avseende vid liknande aspekter.

Analys av relationen till en B1-nivå

Resultatet av analyserna visar att andelen elevtexter som bedöms uppnå en B1-nivå enligt GERS ökar med språksteget. Medelvärdet för bedömningarna på språkstegen *Tyska 3* och *Tyska 4* ligger dock under gränsen för en helt uppnådd nivå B1, medan bedömningarna av elevprestationerna på *Tyska 5* i regel ligger tydligt högre (se kap. 8.1, figur 11). I de aktuella riktlinjerna för ämnet *Moderna språk* anges att den lägsta godkända nivån för det femte

språksteget, alltså *Tyska 5*, motsvaras av en helt uppnådd B1-nivå enligt GERS (B1.2). Analysen av bedömningarna inom *Tyska 5* visar att godkända elevtexter på *Tyska 5* generellt uppnår den förväntade språknivån B1. Dessutom uppfyller även vissa av elevtexterna på de lägre kurserna *Tyska 3* och *Tyska 4* kraven för en helt uppfylld B1-nivå, om än i lägre grad och i relation till de högre betygsstegen. Vidare visar korrelationsberäkningar mellan de svenska bedömningarna och GERS-bedömningarna på en tydlig relation. Korrelationen mellan de svenska bedömningarna och delaspekter i den mer analytiskt inriktade GERS-bedömningen visar vidare att de svenska bedömningarna i högre grad korrelerar med aspekter som har med *formella strukturer* och *ordförråd* än med *fullgörande av uppgiften*.

Diskussion

I centrum för studien står aspekter av validitet vid bedömning av elevers skriftliga kompetens på tyska som kan undersökas i en analys av hur provresultat kan tolkas och användas. Utgångspunkt för diskussionen bildar olika aspekter av validitet enligt Weirs sociokognitiva ramverk för validering (2005) samt relevanta delar av inferenser i en argumentbaserad kedja för validering (jfr Kane 2006; 2013; Chapelle 2020).

Inferens för bedömning och förklaring: konstrukt-konceptualisering

Inferensen för bedömning (*scoring*) innefattar hur en elevprestation omsätts till provresultat som är observerbara och förutsätter adekvata bedömningskriterier (Kane 2013). I en utvidgad version av den argumentbaserade ansatsen återfinns öven inferensen för förklaring (*explanation*), vilken innebär huruvida provresultatet reflekterar det avsedda konstruktet (Chapelle 2020). Inom båda dessa steg i valideringsprocessen spelar bedömares konceptualisering av konstruktet en central roll. I bedömares konstruktkonceptualisering återspeglas aspekter av konstruktvaliditet (*construct validity*), den mest centrala aspekten av validitet enligt Messick (1989a). I Weirs sociokognitiva ramverk (2005) ses konstruktvaliditet som en funktion av interaktionen mellan å ena sidan aspekter av kognitiv validitet (*cognitive validity*) samt kontextvaliditet (*context validity*) och å andra sidan bedömningskriterierna.

Det faktum att resultatet av studien visar att bedömare fäster avseende vid ett brett spektrum av olika aspekter med både negativa och positiva kommentarer tyder på att provformatet dels kan härledas till olika kognitiva färdigheter hos eleverna, dels kan relateras till de olika språkliga och innehållsliga krav

som ställs på testdeltagarna i uppgift och bedömningskriterier. Vidare indikerar det breda spektrumet av aspekter som bedömare fäster avseende vid även att en stor bredd av olika delar av den kommunikativa kompetensen (jfr Bachman & Palmer 1996) beaktas vid bedömningen av skriftlig förmåga. Bedömarnas breda konceptualisering av konstruktet samspelar sålunda med bredden i den teoretiska modellen över kommunikativ språkkompetens och är i linje med tidigare studier (jfr Iwashita m.fl. 2008; Bøhn 2016; Borger 2018). Bedömarna i studien tycks huvudsakligen ha en samstämmig bild av konstruktet, d.v.s. av det som ska prövas, men lyfter emellanåt olika aspekter i texterna och viktar även aspekterna på olika sätt. Dessa olikheter kan förklaras av individuella skillnader mellan enskilda bedömare, något som även återfinns i tidigare studier (jfr Eckes 2008; Borger 2018; Håkansson Ramberg 2021a).

Det går även att urskilja skillnader mellan bedömargrupperna. Dessa skillnader mellan bedömargrupperna kan huvudsakligen härledas till de olika bedömningsskalorna och bedömartraditioner där svenska bedömare bl.a. i högre grad ger ett mer övergripande intryck av texten, både ur ett språkligt och ur ett mer sammanfattande helhetsperspektiv. Vidare visar analysen att svenska bedömare, särskilt de undervisande lärarna, har en tendens att fästa mer avseende vid *lingvistiska aspekter* som språklig korrekthet. Detta är i linje med tidigare studier som visat att bedömare ofta beaktar språkriktighet vid bedömning av prestationer i språk (jfr McNamara 1996; Kuiken & Vedder 2014; Borger 2018). Därutöver skulle vissa framträdande aspekter kunna knytas till andra faktorer, t.ex. kan det jämförelsevis höga antalet kommentarer angående *begriplighet* ha med den språkliga nivån hos eleverna att göra (jfr Pollitt & Murray 1996).

Vidare kan konstateras att svenska bedömare i högre grad fäster avseende vid flera olika aspekter medan GERS-bedömare i sina kommentarer har en mer balanserad fördelning av de aspekter som lyfts fram som relevanta vid en bedömning. Detta tycks till stor del kunna förklaras genom användningen av olika bedömningsskalor, men även genom bedömningsförfarandet, ett mer holistiskt respektive mer analytiskt angreppssätt. Att svenska bedömare däremot ibland lyfter fram olika aspekter i sina bedömningar och i viss mån även viktar och tolkar bedömningskriterier olika är något som kan leda till stora skillnader i betygsättning av elevprestationer. Ett sätt att motverka detta skulle vara att ge lärare ökade möjligheter att sam- och medbedöma elevprestationer för att uppnå en ökad samsyn kring bedömningskriterier samt motverka att enskilda aspekter viktas högre eller på olika sätt av olika bedömare.

Inferens för generalisering: bedömningsvaliditet

Inferensen för generalisering (*generalization*) handlar om det som i skrivbedömningssammanhang ofta benämns *reliabilitet*, vilket bl.a. innebär huruvida olika bedömare genomför en liknande bedömning av samma elevprestation. Inom Weirs sociokognitiva modell (2005) återfinns reliabilitet och samstämmighet mellan olika bedömare i aspekter av bedömningsvaliditet (*scoring validity*). Vad beträffar bedömningsvaliditeten i studien tyder resultaten på att samstämmigheten mellan svenska bedömare vid bedömning av skriftlig språkförmåga inte alltid uppnår en tillfredsställande nivå, vilket kan ha att göra med subjektivitet vid bedömning av uppsatsprov (jfr Bachman m.fl. 1995, Eckes 2011). Skillnader i samstämmighet mellan de svenska bedömarna kan emellertid framför allt observeras gällande konsensusvärdena, något som påvisats även i andra studier vid bedömning av fri textproduktion (jfr Eckes 2011; Tengberg m.fl. 2017). De svenska bedömarna i studien tycks alltså vara mer överens gällande rangordningen i bedömningen än vid en exakt överensstämmelse. Skillnader mellan konsensus- och konsistensvärden förekommer även i andra studier i svensk kontext (jfr Johansson 2013; Tengberg m.fl. 2017) samt vid bedömning av skriftlig förmåga i tyska (t.ex. Bärenfänger 2016).

Att svenska bedömare är mer överens vid bedömning av elevprestationer som erhåller lägre betyg än vid bedömning av elevtexter som erhåller högre betygssteg har även kunnat påvisas i tidigare forskning (jfr Erickson 2009; Granfeldt & Ågren 2014). En möjlig förklaring till att bedömare är mer överens om underkända texter kan vara att dessa relativt tydligt inte uppnår kriteriena för ett godkänt, t.ex. genom att en uppgift saknas. Vidare förekommer även i tidigare studier att de egna undervisande lärarna har en tendens att bedöma sina egna elevers prestationer mildare jämfört med externa bedömare (t.ex. Harlen 2005). Huruvida detta har att göra med att externa bedömare – medvetet eller omedvetet – vid en andra bedömning oftare bedömer strängare (jfr Gustafsson & Erickson 2013) är dock oklart. Här kan även bedömarerfarenhet spela in som faktor då de externa bedömarna var utvalda på grund av sin särskilda erfarenhet i att bedöma elevtexter på tyska. Ytterligare en faktor är att de undervisande lärarna bedömde handskrivna elevtexter, medan de externa bedömarna hade datorskrivna texter till förfogande och att skrivfel inte är lika iögonfallande när de skrivs för hand. En slutsats när det gäller bedömningsvaliditet är vikten av att sambedöma och diskutera bedömning av skriftliga elevprestationer, inte enbart inom den egna skolan, utan även mellan skolor, kommuner och landsdelar. Dessutom är det viktigt att lärare vid sambedömning även fokuserar på elevlösningar som erhåller betyg i mitten eller högre

betyg. Det kan även vara av vikt att undervisande lärare inte alltid bedömer sina egna elevers textproduktion.

Inferens för extrapolering: kriterierelaterad validitet

Inferensen för extrapolering (*extrapolation*) innebär huruvida ett provresultat (här resultatet på ett skriftligt prov på en B1-nivå) kan anses vara en indikator på inlärares språkliga kompetens (här skrivförmåga gällande en uppfylld B1-nivå). Kriterierelaterad validitet (*criterion-related validity*) inom det sociokognitiva ramverket handlar därmed om relationen mellan provresultatet och ett externt kriterium (*criterion*) som antas visa på samma kompetens (Weir 2005). I föreliggande studie fokuseras i detta steg huruvida bedömningar i en svensk skolkontext kan relateras till en extern referensnivå enligt den erkända europeiska referensramen för språk, GERS. Det handlar med andra ord om en mindre empirisk validering av elevers skriftliga kompetens i relation till ett yttre kriterium som ska motsvara samma språkliga kompetensnivå.

Resultatet visar att elevtexter med ett godkänt resultat på *Tyska 5* generellt bedömdes ha uppfyllt kraven för en B1-nivå enligt GERS, d.v.s. att de motsvarande den förväntade språknivå som ställs upp i styrdokumenten. Det faktum att de båda GERS-bedömarna i alla fall utom ett var överens om huruvida en elevlösning hade uppnått en helt uppfylld B1-nivå eller inte stärker studiens resultat. Att ett fåtal elevtexter på *Tyska 5* erhöll både ett godkänt och ett icke godkänt resultat indikerar också att dessa lösningar ligger på gränsen. Resultatet är i linje med tidigare studier av svenska elevers prestationer i engelska på gymnasienivå, där en relation mellan förväntad språknivå enligt GERS och motsvarande språksteg kunnat påvisas (jfr Tyllered 2002; Borger 2018). Däremot står resultatet i kontrast till tidigare studier i moderna språk ur en svensk kontext (jfr European Commission 2012b; Granfeldt m.fl. 2019b; Aronsson 2020). Samtliga av de tidigare studierna i moderna språk fokuserade emellertid på elevprestationer på en lägre språknivå, en A2-nivå enligt GERS. Dessutom har endast en av dessa studier undersökt elevers språkkunskaper i tyska, nämligen det s.k. TAL-projektet som undersöker muntlig kompetens. Sammantaget tycks alltså elevprestationer i språk på gymnasiet i högre grad motsvara kompetensnivåerna enligt GERS i jämförelse med elevprestationer i språk från grundskolan.

Dessutom visar resultatet på att även elevtexter som erhöll ett högre betyg på de lägre kurserna *Tyska 3* och *Tyska 4* uppnår en helt uppfylld B1-nivå enligt GERS. Få tidigare studier har undersökt andelen elevlösningar på lägre språksteg som når högre än den förväntade språknivån. Europarådets ESCL-studie

(jfr European Commission 2012b) visade att enbart några få procent av elevtexterna i spanska uppnådde en högre språknivå i slutet av grundskolan. Huruvida det faktum att även elevlösningar på lägre språksteg uppnår en högre språknivå i tyska skulle kunna förklaras med att eleverna på gymnasienivå är äldre, har fortsatt med sitt valda språk och graden av typologisk likhet mellan svenska och tyska i jämförelse med exempelvis spanska.

Vidare aspekter av kriterierelaterad validitet har undersökts genom korrelationsberäkningar mellan de svenska bedömningarna och resultatet av GERS-bedömningen. Analysen visar på en stark korrelation, vilket indikerar att bedömningarna baseras på liknande konstrukt. Därutöver korrelerades de svenska bedömningarna med aspektbedömningar hos GERS-bedömarna, vilket likaledes visade på en stark relation, särskilt för bedömningsaspekterna *formella strukturer* och *ordförråd*. Detta skulle kunna förklaras av att de svenska bedömarna i studien uppvisar en viss tendens att fästa mer avseende vid lingvistiska aspekter. Även om studien tyder på ett starkt samband mellan svenska bedömningar och en extern referensnivå, vilket stärker den kriterierelaterade validiteten av bedömningen, är materialet i studien relativt begränsat. Sammantaget visar studien på nödvändigheten av en större empirisk validering innehållande ett större material, fler färdigheter än skriftlig kompetens samt även de andra skolspråken som exempelvis spanska och franska.

Slutord

I föreliggande studie undersöktes validitetsaspekter vid bedömning av elevers skriftliga kompetens i tyska inom en svensk utbildningskontext. Här sammanfattas några av de mest centrala slutsatserna. Studien visar att bedömare överlag fäster avseende vid ett brett spektrum av aspekter i den kommunikativa kompetensen vid bedömning av elevtexter på tyska, vilket indikerar att bedömare har en bred konceptualisering av konstruktet. Utmaningen ligger i att i än högre grad nå en gemensam förståelse för vad som ska bedömas och därigenom minska risken för att vissa aspekter ges mer utrymme än andra eller att olika tolkningar görs av bedömningskriterierna.

Vidare visar analysen på utmaningar beträffande bedömningsvaliditeten när det gäller att bedöma elevtexter på ett reliabelt sätt. Bedömarna i studien visar på god förmåga att rangordna elevernas prestationer, men tycks ha svårare att nå exakt överensstämmelse och att förhålla sig till elevprestationer från andra klasser och skolor (jfr Skolverket 2020b). Studien visar även på vikten av att inte enbart diskutera lägstanivån för elevlösningar utan också för elevtexter som erhåller betyg i mitten eller högre betyg. Sam-eller medbedömning

uppmuntras vid prov av den skriftliga förmågan i det frivilliga nationella bedömningsstödet för moderna språk. Det tycks däremot ske i en mindre utsträckning då lärare i moderna språk ofta är ensamma i sitt språkämne på sin skola och upplever att det inte alltid ges tid till detta (jfr Håkansson Ramberg 2021). Då det nationella bedömningsstödet i moderna språk inte är obligatoriskt i det nuvarande systemet och organisationen av sambedömning mellan skolor ofta är bristfällig, kan sådana diskussioner emellertid vara svårare att genomföra. Förhållanden och förutsättningar för betygsättning och bedömning kan dessutom skilja stort mellan olika skolor, vilket i sin tur även kan påverka likvärdigheten vid bedömning. Fortbildningsinsatser och diskussioner om bedömning för redan yrkesverksamma lärare samt under lärarutbildningen skulle i förlängningen kunna bidra till en ökad samsyn för vad som ska bedömas och därmed vara ett stöd för att överbrygga olikheter och skillnader i bedömning mellan skolor, kommuner och landsdelar.

Därutöver pekar studien i riktning mot en hög kriterierelaterad validitet då godkända elevprestationer inom *Tyska 5* i hög grad bedömdes ha uppfyllt kraven för den förväntade referensnivån B1 enligt GERS (B1.2). Detta är en tydlig kvalitetsindikator, särskilt med tanke på att andra tidigare studier från grundskolan visat att elevprestationer i moderna språk inte uppnår förväntad språknivå enligt GERS. Vidare visar resultaten på att elevers språkkunskaper redan på lägre nivåer kan ligga över den förväntade nivån, vilket bör beaktas i undervisningen.

Avslutningsvis visar studien på ett behov av kompetensutveckling för blivande och yrkesverksamma lärare, såväl vad gäller bedömning av elevprestationer i språk som användningen av GERS som referenspunkt. Inom lärarutbildningen och i vidare fortbildningsinsatser för att stärka lärares bedömarkompetens är det därför eftersträvansvärt att svenska bedömare får sätta sig in olika validetsaspekter för bedömning i språk, kunna förhålla sig till innehållet i GERS samt ges möjlighet att reflektera över användning och konsekvenser av bedömningen i relation till den egna språkundervisningen. Detta skulle bidra till lärares professionella utveckling i bedömning och kunna leda till såväl ökad validitet som en högre likvärdighet för elever i olika klassrum på olika skolor vid bedömning av elevers skriftliga kompetens i språk.

Literaturverzeichnis

Abel, A.; Vettori, C. & Wisniewski, K. (Hrsg.) (2012). *KOLIPSI. Die südtiroler SchülerInnen und die Zweitsprache: Eine linguistische und sozialpsychologische Untersuchung*. Bozen: Eurac.

Alderson, J. C. (2005). Editorial. *Language Testing*, 22(3), S. 257–260.

Alderson, J. C. (2007). The CEFR and the need for more research. *The Modern Language Journal* 91, S. 659–663.

Alderson, J. C.; Figueras, N.; Kuijper, H.; Nold, G.; Takala, S. & Tardieu, C (2006). Analysing tests of reading and listening in relation to the Common European Framework of reference: the experience of the Dutch CEFR construct project. *Language Assessment Quaterly* 3(1), S. 3–30.

ALTE (2007). *Minimumstandards for establishing quality profiles in ALTE examinations*. <https://www.alte.org/resources/Documents/minimum_standards_en.pdf> (Abgerufen: September 2019.)

ALTE (2011). *Manual for language test development and examining*. Strasbourg: Council of Europe, Language Policy Division.

American Council on the Teaching of Foreign Languages (2012). *ACTFL proficiency guidelines*. <https://www.actfl.org/resources/actfl-proficiency-guidelines-2012> (Abgerufen: Februar 2017.)

American Psychological Association (1954). *Technical recommendations for psychological tests and diagnostic techniques*. Washington, DC: American Psychological Association.

American Psychological Association, American Educational Research Association & National Council on Measurement in Education (1966). *Standards for educational and psychological tests and manuals*. Washington, DC: American Psychological Association.

American Psychological Association, American Educational Research Association & National Council on Measurement in Education (1974). *Standards for educational and psychological tests*. Washington, DC: American Psychological Association.

American Educational Research Association, American Psychological Association & National Council on Measurement in Education (1985). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

American Educational Research Association, American Psychological Association & National Council on Measurement in Education (1999). *Standards for*

educational and psychological testing. Washington, DC: American Psychological Association.

American Educational Research Association, American Psychological Association & National Council on Measurement in Education (2014). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

Andered, B. (2001). Europarådets Framework – en inspirationskälla för de svenska kursplanerna [Der Referenzrahmen des Europarats – eine Quelle der Inspiration für die schwedischen Lehrpläne]. In: P. Malmberg & R. Ferm (Hrsg.), *Språkboken – en antologi om språkundervisning och språkinlärning.* Stockholm: Skolverket, S. 26–37.

Aronsson, B. (2020). A study of learner profiles in Spanish as a second language in a Swedish instructional setting: Writing versus speaking. *Journal of Linguistics and Language teaching* 11(1), S. 69–90.

Bachman, L. F. (1990). *Fundamental considerations in language testing.* Oxford: University Press.

Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly* 2(1), S. 1–34.

Bachman, L. F.; Lynch, B. K. & Mason, M. (1995). Investigating variability in tasks and rater judgements in a performance test of foreign language speaking. *Language Testing* 12(2), S. 238–257.

Bachman, L. F. & Palmer, A. S. (1996). *Language testing in practice*. Oxford: Oxford University Press.

Bachman, L. F. & Palmer, A. S. (2010). *Language assessment in practice*. Oxford: Oxford University Press.

Bardel, C.; Falk, Y. & Lindqvist, C. (Hrsg.) (2016). *Tredjespråksinlärning* [Drittsprachenerwerb]. Lund: Studentlitteratur.

Bardel, C.; Erickson, G. & Österberg, R. (2019). Learning, teaching and assessment of second foreign languages in Swedish lower secondary school – dilemmas and prospects. *Journal of Applied Language Studies* 13(1), S. 7–26.

Bärenfänger, O. (2016). Die Skalen des Gemeinsamen europäischen Referenzrahmens für Sprachen im Praxistest: Eine empirische Studie zur Validität des Referenzrahmens. *Zeitschrift für Fremdsprachenforschung* 27(1), S. 59–76.

Barkaoui, K. (2007) Rating scale impact on ESL essay marking: A mixed-methods study. *Assessing Writing* 12(2), S. 86–107.

Barkaoui, K. (2010a). Do ESL essay raters' evaluation criteria change with experience? A mixed-methods, cross-sectional study. *TESOL Quarterly* 44(1), S. 31–57.

Barkaoui, K. (2010b). Explaining ESL essay holistic scores: A multilevel modeling approach. *Language Testing* 27(4), S. 515–535.

Barkaoui, K. (2011a). Effects of marking method and rater experience on ESL essay scores and rater performance. *Assessment in Education: Principles, Policy & Practice* 18(3), S. 279–293.

Barkaoui, K. (2011b). Think-aloud protocols in research on essay rating. An empirical study of their veridicality and reactivity. *Language Testing* 28(1), S. 51–75.

Barrett, P. (2001). *Assessing the reliability of rating data*. <https://www.pbarrett.net/presentations/rater.pdf> (Abgerufen: August 2020.)

Berge, K. L. (2005). Skrivprøvenes pålitelighet [Zuverlässigkeit der Schreibtests]. In: K. L. Berge, L. S. Evensen, F. Hertzberg & W. Vagle (Hrsg.), *Ungdommers skrivekompetanse. Bind I: Norsksensuren som kvalitetsutvurdering.* Oslo: Universitetsforlaget, S. 101–113.

Bernhardsson, P. (2016). *I privat och offentligt. Undervisningen i moderna språk i Stockholm 1800–1880* [Im Privaten und im Öffentlichen. Unterricht moderner Sprachen in Stockholm 1800–1880]. Uppsala: Acta Universitatis Upsaliensis.

BIFIE (Hrsg.) (2012). *Bildungsstandards in Österreich. Überprüfung und Rückmeldung* (4. aktualisierte Aufl.). Salzburg: BIFIE. <https://www.bifie.at/wp-content/uploads/2017/06/BIST_Rueckmeldung_Broschuere_web_uk_100812.pdf> (Abgerufen: März 2020.)

Birkel, P. & Birkel, C. (2002). Wie einig sind sich Lehrer bei der Aufsatzbeurteilung? Eine Replikationsstudie zur Untersuchung von Rudolf Weiss. *Psychologie in Erziehung und Unterricht* 49, S. 219–224.

Bøhn, H. (2016). *What is to be assessed? Teachers' understanding of constructs in an oral English examination in Norway.* [Diss.] Oslo: University of Oslo.

Borger, L. (2018). *Investigating and validating spoken interactional competence: Rater perspectives on a Swedish national test of English*. [Diss.] Göteborg: Acta Universitatis Gothoburgensis.

Borsboom, D.; Mellenbergh, G. & van Heerden, J. (2004). The concept of validity. *Psychological Review* 111(4), S. 1061–1071.

Bortz, J. & Döring, N. (2002). *Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler* (3. Aufl.). Berlin: Springer.

Brenner, H. & Kliebsch, U. (2009). Dependence of weighted kappa coefficients on the number of categories. *Epidemiology* 7(2), S. 199–202.

Bringer, J. D.; Johnston, L. H. & Brackenridge, C. H. (2004). Maximizing transparency in a doctoral thesis: the complexities of writing about the use of QSR*NVIVO within a grounded theory. *Qualitative Research* 4(2), S. 245–265.

Broek, S. & van den Ende, I. (2013). *Die Umsetzung des Gemeinsamen europäischen Referenzrahmens für Sprachen in den europäischen Bildungssystemen*. Brüssel: Europäisches Parlament, Fachabteilung B: Struktur- und Kohäsionspolitik. <https://www.europarl.europa.eu/RegData/etudes/etudes/join/2013/495871/IPOL-CULT_ET%282013%29495871_DE.pdf> (Abgerufen: März 2019.)

Brown, A.; Iwashita, N. & McNamara, T. (2005). An examination of rater orientation and test-taker performance on English-for-academic-purposes speaking tasks. *TOEFL-MS-29*. Princeton, NJ: Educational Testing Service.

Cabau-Lampa, B. (2005). Foreign language education in Sweden from a historical perspective: Status, role and organization. *Journal of Educational Administration and History* 37(2), S. 91–111.

Cabau-Lampa, B. (2007). Mother tongue plus two European languages in Sweden: unrealistic educational goal? *Language Policy* 6, S. 333–358.

Canale, M. & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics* 1(1), S. 1–47.

Canale, M. (1983). From communicative competence to communicative language pedagogy. In: J. C. Richards & R. W. Schmidt (Hrsg.), *Language and communication*. London: Longman, S. 2–27.

Cardelús, E. (2015). *Motivationer, attityder och moderna språk. En studie om elevers motivationsprocesser och attityder vid studier och lärande av moderna språk* [Motivationen, Einstellungen und *Moderna språk*. Eine Studie über Motivationsprozesse und Einstellungen von Schülerinnen und Schülern beim Erlernen einer modernen Sprache]. [Diss.] Stockholms universitet: Institutionen för språkdidaktik.

Centre for Canadian Language Benchmarks (2019). *The Canadian language benchmarks*. <https://www.language.ca/overview-of-clb-and-nclc-competency-levels/> (Abgerufen: November 2020.)

Chalhoub-Deville, M. (1997). Theoretical models, assessment frameworks and test construction. *Language Testing* 14(1), S. 3–22.

Chapelle, C. A. (1998). Construct definition and validity inquiry in SLA research. In: L. F. Bachman & A. D. Cohen (Hrsg.), *Interfaces between second language acquisition and language testing research*. New York/Cambridge: Cambridge University Press, S. 32–70.

Chapelle, C. A. (2020). *Argument-based validation in testing and assessment*. Los Angeles: SAGE.

Chapelle, C. A.; Enright, M. K. & Jamieson, J. (2008). Building a validity argument for the test of English as a foreign language. New York: Routledge.

Chapelle, C. A.; Enright, M. K. & Jamieson, J. (2010). Does an argument-based approach to validity make a difference? *Educational Measurement: Issues and Practice* 29(1), S. 3–13.

Chapelle, C. A. & Voss, E. (2013). Evaluation of language tests through validation research. In: A. J. Kunnan (Hrsg.), *The companion to language assessment*. Boston: Wiley-Blackwell, S. 1081–1097.

Chapelle, C.; Kremmel, B. & Brindley, G. (2020). Assessment. In: N. Schmitt & M. P. H. Rodgers (Hrsg.), *An introduction to applied linguistics*. Abingdon, Oxon: Routledge, S. 294–316.

Cohen, J. (1968). Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin* 70(4), S. 213–220.

Connolly, S.; Klenowski, V. & Wyatt-Smith, C. M. (2012). Moderation and consistency of teacher judgement: teachers' views. *British Educational Research Journal* 38(4), S. 593–614.

Holec, H.; Little, D. & Richterich, R. (Hrsg.) (1996). *Strategies in language learning and use: Studies towards a Common European Framework of reference for language learning and teaching*. Strasbourg: Council of Europe Publishing.

Council of Europe (2008). *Recommendation CM/Rec (2008)7 of the Committee of Ministers to member states on the use of the Council of Europe's Common European Framework of Reference for Languages (CEFR) and the proportion of plurilingualism*. <https://www.ecml.at/Portals/1/documents/CoE-documents/Rec-CM-2008-7-EN.pdf?ver=2016-11-29-112711-910> (Abgerufen: Oktober 2019.)

Council of Europe (2009). *Relating language examinations to the Common European framework of reference for languages. Learning, teaching, assessment (CEFR). A Manual*. Strasbourg: Language Policy Division.

Council of Europe (2020). *Gemeinsamer europäischer Referenzrahmen für Sprachen: lernen, lehren, beurteilen. Begleitband*. Stuttgart: Klett Sprachen GmbH/Straßburg: Council of Europe.

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, Mass.: MIT Press.

Cronbach, L. J. & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin* 52(4), S. 281–302.

Cronbach. L. J. (1971). Test validation. In: R. L. Thorndike (Hrsg.), *Educational Measurement*. Washington, D. C.: American Council on Education, S. 443–507.

Crooks, T. J.; Kane, M. T. & Cohen, A. S. (1996). Threats to the valid use of assessments. *Assessment in Education* 3(3), S. 265–285.

Cumming, A. (1990). Expertise in evaluating second language compositions. *Language Testing* 7(1), S. 31–51.

Cumming, A.; Kantor, R. & Powers, D. E. (2002). Decision making while rating ESL/EFL writing tasks: A descriptive framework. *The Modern Language Journal* 86, S. 67–96.

Cureton, E. E. (1951). Validity. In: E. F. Lindquist (Hrsg.), *Educational Measurement*. Washington, D. C.: American Council on Education, S. 621–694.

Dalberg, T. (2019). *Samstämmighet i skrivbedömning. Statistisk analys vid bedömning av två nationella skrivprov* [Übereinstimmung bei der Bewertung schriftlicher Leistungen. Statistische Analyse für die Bewertung zweier nationaler Tests des schriftlichen Ausdrucks]. (Svenska i utveckling 36.) Uppsala: Uppsala universitet.

Davis, L. (2016). The influence of training and experience on rater performance in scoring spoken language. *Language Testing*, 33(1), S. 117–135.

De Florio Hansen, I. (2015). *Standards, Kompetenzen und fremdsprachliche Bildung. Beispiele für den Englisch- und Französischunterricht*. Tübingen: Narr Francke Attempto Verlag Gmbh.

DESI-Konsortium (2006). *Unterricht und Kompetenzerwerb in Deutsch und Englisch. Zentrale Befunde der Studie Deutsch-Englisch-Schülerleistungen-International (DESI)*. Frankfurt am Main: Deutsches Institut für Internationale Pädagogische Forschung (DIPF).

Douglas, D. (2010). *Understanding language testing*. London: Hodder Education.

Ducasse, A. M. & Brown, A. (2009). Assessing paired orals: Raters' orientation to interaction. *Language Testing* 26(3), S. 423–443.

EALTA (2006). *Guidelines for good practice in language testing and assessment*. <http://www.ealta.eu.org/documents/archive/guidelines/English.pdf> (Abgerufen: September 2019.)

Eckes, T. (2004). Facetten des Sprachtestens: Strenge und Konsistenz in der Beurteilung sprachlicher Leistungen. In: A. Wolff, T., A. Ostermann & C. Chlosta (Hrsg.), *Integration durch Sprache*. Regensburg: FaDaF, S. 485–518.

Eckes, T. (2005). Examining rater effects in TestDaF writing and speaking performance assessments: A Many Facet Rasch analysis. *Language Assessment Quarterly* 28(3), S. 197–221.

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing* 25(2), S. 155–185.

Eckes, T. (2011). Facetten der Genauigkeit. Zur Reliabilität der Beurteilung fremdsprachlicher Leistungen. *Deutsch als Fremdsprache* 48(4), S. 195–204.

Eckes, T. (2015). *Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments*. Frankfurt am Main: Peter Lang.

Eckes, T. (2019). Many-facet Rasch measurement. Implications for rater-mediated language assessment. In: V. Aryadoust & M. Raquel (Hrsg.), *Quantitative data analysis for language assessment. Volume I. Fundamental techniques*. London/New York: Routledge, S. 153–175.

Eckes, T.; Müller-Karabil, A. & Zimmermann, S. (2016). Assessing writing. In: D. Tsagari, & J. Banerjee (Hrsg.), *Handbook of second language assessment*. Boston/Berlin: De Gruyter Mouton, S. 147–164.

Erickson, G. (2009). Nationella prov i engelska – en studie av bedömarsamstämmighet [Nationale Tests im Fach Englisch – eine Studie der Beurteilerübereinstimmung]. <https://nafs.gu.se/digitalAssets/1319/1319572_bed.np- eng-g.erickson-2009.pdf> (Abgerufen: Januar 2020.)

Erickson, G. (2011a). Handle with care. Om referensramen och bedömning av språklig kompetens [Über den Referenzrahmen und die Bewertung sprachlicher Kompetenz]. In: Söderberg, C. (Hrsg.), *Språklärarens stora blå? En samling texter om Gemensam europeisk referensram för språk*. Uppsala universitet: Fortbildningsavdelningen för skolans internationalisering, S. 31–37.

Erickson, G. (2011b). Putting the CEFR to good use – A collaborative challenge. In: J. Mader & Z. Urkun (Hrsg.), *Putting the CEFR to good use – IATEFL TEA SIG/EALTA Conference Proceedings, Barcelona 2010*, S. 36–43. <http://www.ealta.eu.org/documents/resources/IATEFL_EALTA_Proceedings_2010.pdf> (Abgerufen: Februar 2020.)

Erickson, G. (2019). Holistic peer analyses of national tests in relation to the CEFR. In: A. Huhta, G. Erickson & N. Figueras (Hrsg.), *Developments in language education: A memorial volume in honour of Sauli Takala*. Juväskylä: EALTA & University of Juväskylä, S. 49–66.

Erickson, G. (2020a). Finding out what learners know – and …? Reflections on teachers' language assessment literacy. In: D. Tsagari (Hrsg.), *Language assessment literacy: From theory to practice*. Cambridge: Cambridge Scholars Publishing, S. 29–47.

Erickson, G. (2020b). *National assessment of foreign languages in Sweden*. <https://www.gu.se/sites/default/files/2020-04/Nat_Assesment_of_Foreign_Lang_in_Swe2020.pdf> (Abgerufen: April 2020.)

Erickson, G. & Åberg-Bengtsson, L. (2012). A collaborative approach to national test development. In: D. Tsagari & I. Csépes (Hrsg.), *Collaboration in language testing and assessment*. Frankfurt am Main: Peter Lang, S. 93–108.

Erickson, G. & Pakula, H.-M. (2017). Den gemensamma europeiska referensramen för språk: Lärande, undervisning, bedömning – ett nordiskt perspektiv

[Der gemeinsame europäische Referenzrahmen für Sprachen: lernen, lehren, beurteilen – eine nordische Perspektive]. *Acta Didactica Norge* 11(3), S. 1–23.

Erickson, G. & Sylvén, L. K. (2013). Lärande och bedömning i språk [Das Lernen und die Bewertung von Sprachen]. In: I. Wernersson & I. Gerrbo (Hrsg.), *Differentieringens janusansikte: en antologi från Institutionen för pedagogik och specialpedagogik vid Göteborgs* universitet. Göteborg: Acta Universitatis Gothoburgensis, S. 77–114.

Erickson, G.; Österberg, R. & Bardel, C. (2018). Lärares synpunkter på ämnet Moderna språk – en rapport från projektet TAL [Ansichten der Lehrkräfte im Fach *Moderna språk* – ein Bericht aus dem Projekt TAL]. *LMS – Lingua 2*, S. 8–12.

Europäische Union (2014). *Schlussfolgerungen des Rates vom 20. Mai 2014 zur Mehrsprachigkeit und zur Entwicklung von Sprachenkompetenz.* <https://eur-lex.europa.eu/legal-content/DE/TXT/?uri=uriserv:OJ.C_.2014.183.01.0026.01.DEU> (Abgerufen: September 2021.)

Europarat (2001). *Gemeinsamer europäischer Referenzrahmen für Sprachen: lernen, lehren, beurteilen*. München: Langenscheidt/Straßburg: Europarat.

European Commission (2012a). *Eurobarometer 386. Europeans and their languages.* <https://ec.europa.eu/commfrontoffice/publicopinion/archives/ebs/ebs_386_en.pdf> (Abgerufen: Oktober 2019.)

European Commission (2012b). *First European survey on language competences: Final report.* Brussels: European Commission. <https://www.researchgate.net/publication/262877352_First_European_Survey_on_Language_Competences_Final_Report> (Abgerufen: Oktober 2019.)

European Council (2002). Presidency conclusions. Barcelona European Council 15 and 16 March 2002. <https://ec.europa.eu/commission/presscorner/detail/en/PRES_02_930> (Abgerufen: August 2021.)

Fan, J. & Bond, T. (2019). Applying Rasch measurement in language assessment. Unidimensionality and local independence. In: V. Aryadoust & M. Raquel (Hrsg.), *Quantitative data analysis for language assessment Volume I. Fundamental techniques*. London/New York: Routledge, S. 83–102.

Figueras, N. (2009). The impact of the CEFR. *ELT Journal* 66(4), S. 477–485.

Fulcher, G. (2004). Deluded by artifices? The Common European Framework and harmonization. *Language Assessment Quarterly* 1(4), S. 253–266.

Fulcher, G. & Davidson, F. (2007). *Language testing and assessment. An advanced resource book*. London and New York: Routledge.

Fulcher, G. & Davidson, F. (2009). Test architecture, test retrofit. *Language Testing* 26(1), S. 123–144.

Fulcher, G. (2016). Standards and frameworks. In: D. Tsagari & J. Banerjee (Hrsg.), *Handbook of second language assessment*. Boston/Berlin: De Gruyter Mouton, S. 29–44.

Gibbons, S. & Marshall, B. (2010). Assessing English: A trial collaborative standardised marking project. *English Teaching: Practice and Critique* 9(3), S. 26–39.

Gipps, C. V. (1994). Beyond testing: towards a theory of educational assessment. London: The Falmer Press.

Glaboniat, M.; Perlmann-Balme, M. & Studer, T. (2013). *Zertifikat B1. Deutschprüfung für Jugendliche und Erwachsene. Prüfungsziele, Testbeschreibung.* Ismaning: Hueber Verlag.

Goertler, S.; Kraemer, A. & Schenker, T. (2018). Setting evidence-based language goals. *Foreign Language Annals* 49(3), S. 434–454.

Goethe-Institut (2017). *Materialien zur Prüfung Goethe-Zertifikat B1. Übungssatz Jugendliche*. (2. Aufl.) München: Goethe-Institut. <https://www.goethe.de/pro/relaunch/prf/materialien/B1/B1_Uebungssatz_Jugendliche.pdf> (Abgerufen: September 2021.)

Goethe-Institut (2018). *Goethe-Zertifikat B1. Durchführungsbestimmungen*. <http://www.goethe.de/lrn/prf/pro/sv/Durchfuehrungsbestimmungen_B1.pdf> (Abgerufen: Oktober 2018.)

Goethe-Institut (2019). *Deutschprüfungen*. <https://www.goethe.de/de/spr/kup/prf.html> (Abgerufen: Januar 2019.)

Graham, S.; Harris, K. & Herbert, M. (2011). *Informing writing: The benefits of formative assessment*. A Carnegie Corporation Time to Act report. Washington, D. C.: Alliance for Excellent Education.

Granfeldt, J. & Ågren, M. (2014). SLA developmental stages and teachers' assessment of written French: Exploring Direkt Profil as a diagnostic assessment tool. *Language Testing* 31(3), S. 285–305.

Granfeldt, J.; Sayehli, S. & Ågren, M. (2019a). The context of second foreign languages in Swedish secondary schools: Results of a questionnaire to schools leaders. *Apples – Journal of Applied Language Studies* 13(1), S. 27–48.

Granfeldt, J.; Bardel, C.; Erickson, G.; Sayehli, S.; Ågren, M. & Österberg, R. (2019b). Muntlig språkfärdighet i främmande språk – en studie av samspelet mellan lärande, undervisning och bedömning [Mündliche Sprachfähigkeit in einer Fremdsprache – eine Studie zum Zusammenspiel zwischen Lernen, Lehren und Beurteilen]. In: *Vetenskapsrådets Resultatdialog 2019*. Stockholm: Vetenskapsrådet, S. 29–33.

Granfeldt, J.; Sayehli, S. & Ågren, M. (2021). Trends in the study of modern languages in Swedish lower secondary school (2000–2018) and the impact of grade point average enhancement credits. *Education Inquiry*, S. 127–146.

Green, A. (1998). *Verbal protocol analysis in language testing research: A handbook. Studies in Language Testing 5.* Cambridge: Cambridge University Press.

Green, A. (2018). Linking tests of English for academic purposes to the CEFR: The score user's perspective. *Language Assessment Quarterly* 15(1), S. 59–74.

Grotjahn R. (2017): Testkonstrukt und Testspezifikationen. In: B. Akukwe, R. Grotjahn & S. Schipolowski (Hrsg.). *Schreibkompetenzen in der Fremdsprache*. Tübingen: Narr Francke Attempto Verlag Gmbh, S. 71–116.

Gustafsson, J-E. & Erickson, G. (2013). To trust or not to trust? Teacher marking versus external marking of national tests. *Educational Assessment, Evaluation and Accountability* 25(1), S. 69–87.

Gustafsson, J-E.; Cliffordson, C. & Erickson, G. (2014). *Likvärdig kunskapsbedömning i och av den svenska skolan: problem och möjligheter* [Gleichwertige Bewertung von Wissen in und von der schwedischen Schule: Probleme und Möglichkeiten]. Stockholm: SNS förlag.

Håkansson Ramberg, M. (2016). *Was bewerten Lehrer? Die Bedeutung grammatischer und lexikalischer Faktoren bei der Benotung von Schülertexten im Fach Deutsch als Fremdsprache* [Lizentiatarbeit]. Växjö: Linnéuniversitetet.

Håkansson Ramberg, M. (2021). „Det ska vara begripligt" – Om lärares bedömning av godkänd nivå i tyska [„Es sollte verständlich sein" – Über die Bewertung eines ausreichenden Sprachniveaus in Deutsch durch Lehrkräfte]. In: C. Bardel, G. Erickson, J. Granfeldt & C. Rosén (Hrsg.), *Forskarskolan FRAM: Lärare forskar i de främmande språkens didaktik*. Stockholm: Stockholm University Press, S. 177–200.

Hambleton, R. K.; Swaminathan, H. & Rogers, J. H. (1991). *Fundamentals of item response theory.* Newbury Park: SAGE.

Hambleton, R. K.; Jaeger, R. M.; Koretz, D.; Linn, R. L.; Millman, J. & Philips, S. E. (1995). *Review of the measurement quality of the Kentucky Instructional Results Information System, 1991–1994*. Frankfort: Office of Education Accountability, Kentucky General Assembly.

Harding, L. (2014). Communicative Language testing: Current issues and future research. *Language Assessment Quaterley* 11(2), S. 186–197.

Harlen, W. (2005). Trusting teachers' judgement: research evidence of the reliability and validity of teachers' assessment used for summative purposes. *Research Papers in Education* 20(3), S. 245–270.

Harsch, C. (2006). *Der gemeinsame europäische Referenzrahmen: Leistung und Grenzen. Die Bedeutung des Referenzrahmens im Kontext der Beurteilung von Sprachvermögen am Beispiel des semikreativen Schreibens im DESI-Projekt*. [Diss.] Universität Augsburg: Philologisch-Historische Fakultät.

Harsch, C. & Rupp, A. A. (2011). Designing and scaling level-specific writing tasks in alignment with the CEFR: A test-centerd approach. *Language Assessment Quarterly* 8(1), S. 1–34.

Harsch, C. & Martin, G. (2013). Comparing holistic and analytic scoring methods: Issues of validity and reliability. *Assessment in Education: Principles, Policy and Practice* 20(3), S. 281–307.

Harsch, C. & Hartig, J. (2015). What are we aligning tests to when we report test alignment to the CEFR? *Language Assessment Quarterly* 12(4), S. 333–362.

Hildén, R. & Takala, S. (2007). Relating descriptors of the Finnish school scale to the CEF overall scales for communicative activities. In: A. Koskensalo, J. Smeds, P. Kaikkonen, & V. Kohonen (Hrsg.), *Foreign languages and multicultural perspectives in the European context / Fremdsprachen und multikulturelle Perspektiven im europäischen Kontext*. Berlin: LIT-Verlag, S. 291–300.

Hildén, R. (2008). *Analys av svenska kursplaner i relation till den europeiska referensramen* [Analyse der schwedischen Lehrpläne in Relation zum europäischen Referenzrahmen]. [Unpubliziert.]

Hildén, R.; Härmälää, M.; Rautopuro, J. & Huhtanen, M. (2019). Finnisch 9th graders' language skills: Effects of learning environment and teaching on levels attained compared with other European countries. In: A. Huhta, G. Erickson, & N. Figueras (Hrsg.), *Developments in language education: A memorial volume in honour of Sauli Takala*. Juväskylä: EALTA & University of Juväskylä, S. 113–130.

Hsieh, C.-N. (2011). *Rater effects in ITA testing: ESL teachers' versus American undergraduates' judgements of accentedness, comprehensibility, and oral proficiency*. [Diss.] Ann Arbor: Michigan State University.

Hulstijn, J. H. (2007). The shaky ground beneath the CEFR: Quantitative and qualitative dimensions of language proficiency. *The Modern Language Journal* 91(4), S. 663–667.

Hymes, D. (1972). On communicative competence. In: J. B. Pride & J. Holmes (Hrsg.), *Sociolinguistics: selected readings*. Harmondsworth: Penguin, S. 269–293.

Iwashita, N.; Brown, A.; McNamara, T. & O'Hagan, S. (2008). Assessed levels of second language speaking proficiency: How distinct? *Applied Linguistics* 29(1), S. 24–49.

Jang, E. E.; Wagner, M. & Park, G. (2014). Mixed methods research in language testing and assessment. *Annual Review of Applied Linguistics* 34, S. 123–153.

Johansson, S. (2013). *On the validity of reading assessments: relationships between teacher judgements, external tests and pupil self-assessments*. Göteborg: Acta Universitatis Gothoburgensis.

Johansson, S. (2015). Validitet och lärares bedömningar [Validität und Bewertungen der Lehrkräfte]. *Pedagogisk forskning i Sverige* 20(1–2), S. 33–53.

Johnson, R. B. & Onwuegbuzie, A. J. (2004). Mixed methods research: A research paradigm whose time has come. *Educational Researcher* 33(7), S. 14–26.

Jølle, L. (2015). Rater strategies for reaching agreement on pupil text quality. *Assessment in Education: Principles, Policy & Practice* 22(4), S. 458–474.

Jones, N. & Saville, N. (2016). *Learning orientied assessment. A systemic approach*. Cambridge: Cambridge University Press.

Jönsson, A. & Svingby, G. (2007). The use of scoring rubrics: reliability, validity and educational consequences. *Educational Research Review* 2, S. 130–144.

Jönsson, A. & Balan, A. (2018). Analytic or holistic: A study of agreement between different grading models. *Practical Assessment, Research & Evaluation* 23(18), S. 1–11.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin* 112(3), S. 527–535.

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement* 38(4), S. 319–342.

Kane, M. T. (2002). Validating high-stakes testing programs. *Educational Measurement: Issues and Practice* 21(1), S. 31–41.

Kane, M. T. (2006). Validation. In: R. L. Brennan (Hrsg.), *Educational measurement* (4. Aufl.). Westport, CT: American Council on Education/Praeger Publishers, S. 17–64.

Kane, M. T. (2011). Validating score interpretations and uses. *Language Testing* 29(1), S. 3–17.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of educational measurement* 50(1), S. 1–73.

Kane, M.; Crooks, T. & Cohen, A. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice* 18(2), S. 5–17.

Kantarcioglu, E. (2012). Relating an institutional proficiency examination to the CEFR: a case study. [Diss.] London: University of Roehampton. <https://pure.roehampton.ac.uk/ws/portalfiles/portal/443838/Elif_Kantarcioglu_PhD_2012.pdf> (Abgerufen: März 2021.)

Kecker, G. (2011). *Validierung von Sprachprüfungen. Die Zuordnung des TestDaF zum Gemeinsamen europäischen Referenzrahmen für Sprachen*. Frankfurt am Main: Peter Lang.

Kecker, G. (2014). Aktuelle Entwicklungen in der Messung von Sprachkompetenzen: Einführung. *Zeitschrift für Interkulturellen Fremdsprachenunterricht* 19(2), S. 1–4.

Kelley, T. L. (1927). *Interpretation of educational measurements*. New York: Macmillan.

Kendall, M. & Dickinson Gibbons, J. (1990). Rank correlation methods (5. Aufl.). London: Edward Arnold.

Kim, Y-H. (2009). An investigation into native and non-native teachers' judgements of oral English performance: A mixed methods approach. *Language Testing* 26(2), S. 187–217.

Klapp Lekholm, A. (2008). *Grades and grade assignment: effects of student and school characteristics.* [Diss.] Göteborg: Acta Universitatis Gothoburgensis.

Klieme, E. (2006). *Zusammenfassung zentraler Ergebnisse der DESI-Studie*. Frankfurt am Main: Deutsches Institut für Internationale Pädagogische Forschung (DIPF).

Klieme, E.; Avenarius, H.; Blum, W.; Döbrich, P.; Gruber, H.; Prenzel, M.; Reiss, K.; Riquarts, K.; Rost, J.; Tenorth, H. & Vollmer, H. J. (2003). *Zur Entwicklung nationaler Bildungsstandards. Eine Expertise*. Bonn: BMBF.

Klieme, E. & Leutner, D. (2006). Kompetenzmodelle zur Erfassung individueller Lernergebnisse und zur Bilanzierung von Bildungsprozessen. Beschreibung eines neu eingerichteten Schwerpunktprogramms der DFG. *Zeitschrift für Pädagogik* 52(6), S. 876–903.

Knoch, U. & Chapelle, C. A. (2018). Validation of rating processes within an argument-based framework. *Language Testing* 35(4), S. 477–499.

Köller, Ö.; Knigge, M. & Tesch, B. (Hrsg.) (2010). *Sprachliche Kompetenzen im Ländervergleich*. Münster: Waxmann.

Koretz, D. (2008) *Measuring up. What educational testing really tells us*. Cambridge, MA: Harvard University Press.

Krigh, J. (2019) *Språkstudier som utbildningsstrategi hos grundskoleelver och deras familjer* [Sprachenlernen als Bildungsstrategie bei Schülerinnen und Schülern in der Grundschule und ihren Familien]. [Diss.] Uppsala: Acta Universitatis Upsaliensis.

Kuckartz, U. (2014a). *Mixed Methods: Methodologie, Forschungsdesigns und Analyseverfahren*. Wiesbaden: Springer.

Kuckartz, U. (2014b). *Qualitative text analysis: a guide to methods, practice and using software* [Englische Übersetzung]. Los Angeles/London/New Delhi/Singapore/Washington DC: Sage.

Kuiken, F. & Vedder, I. (2014). Rating written performance: What do raters do and why? *Language Testing* 31(3), S. 329–348.

Kunnan, A. J. (2004) Test fairness. In: M. Milanovic & C. J. Weir (Hrsg.), *European language testing in a global context*. Cambridge: Cambridge University Press, S. 27–48.

Landis, J. R. & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics* 33(1), S. 159–174.

Lenz, P. (2006). Überlegungen zur Sprachkompetenzbeschreibung und Testvalidierung im Projekt HarmoS/Fremdsprachen. *Bulletin suisse de linguistique appliquée* 84, S. 191–227.

Lenz, P. & Studer, T. (2008). Zur Entwicklung der Expertenvorschläge für Basisstandards in den Fremdsprachenfächern. *Beiträge zur Lehrerinnen- und Lehrerbildung* 26(3), S. 361–371.

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions* 16(2), S. 878.

Linacre, J. M. (2005). *Facets Rasch measurement computer program*. Version 3.58.0. Chicago: Winsteps.com.

Lissitz, R. W. & Samuelson, K. (2007). A suggested change in terminology and emphasis regarding validity and education. *Educational Researcher* 36(8), S. 437–448.

Lumley, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to the raters? *Language Testing* 19(3), S. 246–276.

Lumley, T. (2005). *Assessing second language writing. The rater's perspective*. Frankfurt am Main: Peter Lang.

Magnan, S. (1988). Grammar and the ACTFL oral proficiency interview: Discussion and data. *The Modern Language Journal* 72(3), S. 266–276.

Malmberg, P. (1986). Språkundervisningen i Sverige i ett historiskt perspektiv [Sprachunterricht in Schweden aus einer historischen Perspektive]. In: B. Andered & T. Lindblad (Hrsg.), *Dagens språkundervisning och morgondagens. Rapport från en konferens om ungdomsskolans språkundervisning i Tällberg den 27–30 januari 1986*. Stockholm: Skolöverstyrelsen, S. 11–21.

May, L. (2011). *Interaction in a paired speaking test. The rater's perspective*. Frankfurt am Main: Peter Lang.

Mayring, P. (2015). *Qualitative Inhaltsanalyse. Grundlagen und Techniken*. (12. überarbeitete Aufl. 2015). Weinheim und Basel: Beltz Verlag.

McKinstry, B. H.; Cameron, H. S.; Elton, R. A. & Riley S. C. (2004) Leniency and halo effects in marking undergraduate short research projects. *BMC Medical Education* 4(28), S. 1–5.

McNamara, T. (1990). Item response theory and the validation of an ESP test for health professionals. *Language Testing* 7(1), S. 52–75.

McNamara, T. (1996). *Measuring second language performance.* New York: Longman.

McNamara, T. (2006). Validity in language testing: The challenge of Sam Messick's legacy. *Language Assessment Quarterly* 3(1), S. 31–51.

McNamara, T. (2010). The use of language tests in the service of policy: issues of validity. *Revue française de linguistique appliquée* 1 (Vol. XV), S. 7–23.

McNamara, T. & Roever (2006). *Language testing: The social dimension*. London: Blackwell.

Mehrens, W. A. (1997). The consequences of consequential validity. *Educational Measurement: Issues and Practice* 16(2), S. 16–18.

Messick, S. (1989a). Meaning and values in test validations: The science and ethics of assessment. *Educational Researcher* 18(2), S. 5–11.

Messick, S. (1989b). Validity. In: R. L. Linn (Hrsg.), *Educational measurement* (3. Aufl.). New York: American Council on Education, S. 13–103.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist* 50, S. 741–749.

Mislevy, R. J.; Steinberg, L. S. & Almond, R. G. (2002). Design and analysis in task-based language assessment. *Language Testing* 19(4), S. 477–494.

Mislevy, R. J. & Riconscente, M. M. (2006). Evidence-centered assessment design. In: S. M. Downing & T. M. Haladyna (Hrsg.), *Handbook of test development*. Mahwah, NJ: Lawrence Erlbaum Associates, S. 61–90.

Molander, P. (2017). *Dags för omprövning – en ESO-rapport om styrning av offentlig verksamhet, Rapport till Expertgruppen för studier i offentlig ekonomi 2017:1* [Zeit für eine erneute Überprüfung – ein ESO-Bericht zur Steuerung von Einrichtungen, Bericht an die Expertengruppe für Studien öffentlicher Wirtschaft]. Stockholm: Regeringskansliet. <https://eso.expertgrupp.se/wp-content/uploads/2017/03/2017_1-Dags-för-omprövning.pdf> (Abgerufen: Dezember 2020.)

Moss, P. A.; Girard, B. J. & Haniford, L. C. (2006). Validity in educational assessment. *Review of Research in Education* 30, S. 109–162.

Murphy, R. (1979). Removing the marks from examination scripts before remarking them: does it make any difference? *British Journal of Educational Psychology* 49(1), S. 73–78.

Newton, P. E. & Shaw, S. (2014). *Validity in educational and psychological assessment*. London: SAGE Publications.

North, B. (2007). *The CEFR Common Reference levels: validated reference points and local strategies*. Strasbourg: Council of Europe, Language Policy Division. <https://rm.coe.int/16805c3896> (Abgerufen: Oktober 2019.)

North, B. (2014). *The CEFR in practice*. Cambridge: Cambridge University Press.

North, B. & Piccardo, E. (2018*). Aligning the Canadian Language Benchmarks (CLB) to the Common European framework of reference (CEFR).* Research report. Toronto: Centre for Canadian Language Benchmarks.

Nusche, D.; Halász, G.; Looney, J.; Santiago, P. & Shewbridge, C. (2011). *OECD reviews of evaluation and assessment in education: Sweden*. Paris: OECD. <https://www.oecd.org/sweden/47169533.pdf> (Abgerufen: Oktober 2019.)

Nyström, P. (2004). *Rätt mätt på prov: om validering av bedömningar i skolan* [engl. Titel: Validation of educational assessments]. Umeå: Umeå universitet.

Oscarson, M. (2015). Bedömning på systemnivå – En komparativ studie av stegsystemet i språk i den svenska skolan och språknivåer i Europarådets Common European Framework of Reference for Languages (CEFR) [Bewertung auf Systemebene – Eine komparative Studie des Stufensystems in der schwedischen Schule und der Sprachniveaus im Gemeinsamen europäischen Referenzrahmen für Sprachen (GER) des Europarats]. In: *EDUCARE*, 2015(2), S. 128–153.

Östlund-Stjärnegårdh, E. (2002). *Godkänd i svenska? Bedömning och analys av gymnasieskolans texter* [Eine ausreichende Note in Schwedisch? Beurteilung und Analyse der Texte am Gymnasium]. (Skrifter utgivna av Institutionen för nordiska språk vid Uppsala universitet 57.) Uppsala: Uppsala universitet.

O'Sullivan, B. (2008*). City and Guilds communicator IESOL examination (B2) CEFR linking project. Case study report*. London: City and Guilds.

O'Sullivan, B. & Weir, C. J. (2011). Test development and validation. In: B. O'Sullivan (Hrsg.), *Language testing: Theories and practices*. Basingstoke: Palgrave Macmillan, S. 13–32.

Papageorgiou, S. (2007). *Relating the trinity College London GESE and ISE examinations to the Common European Framework of Reference. Final project report*. London: Trinity College London.

Papageorgiou, S. (2010). Investigating the decision-making process of standard setting participants. *Language Testing* 27(2), S. 261–282.

Papageorgiou, S. (2016). Aligning language assessments to standards and frameworks. In: D. Tsagari, & J. Banerjee (Hrsg.), *Handbook of second language assessment*. Boston/Berlin: De Gruyter Mouton, S. 327–340.

Papageorgiou, S.; Tannenbaum, R. J.; Bridgeman, B. & Cho, Y. (2015). *The association between TOEFL iBT test scores and the Common European Framework of Reference (CEFR) levels* (Research Memorandum No. RM 15-06). Princeton, NJ: Educational testing Service.

Pollitt, A. & Murray, N. L. (1996). What raters really pay attention to. In: M. Milanovic & N. Saville (Hrsg.). *Performance testing, cognition and assessment*. Cambridge: UCLES/Cambridge University Press, S. 74–91.

Popham, W. J. (1997). Consequential validity: Right concern – wrong concept. *Educational Measurement: Issues and Practice* 16(2), S. 9–13.

Powers, D. E.; Fowles, M. E.; Farnum, M. & Ramsey, P. (1994). Will they think less of my handwritten essay if others word process theirs? Effects on essay scores of intermingling handwritten and word-processed essays. *Journal of Educational Measurement* 31(3), S. 220–233.

Purpura, J. E. (2016). Second and foreign language assessment. *The Modern Language Journal* 100(1), S. 190–208.

Quetz, J. & Vogt, K. (2009). Bildungsstandards für die erste Fremdsprache: Sprachenpolitik auf unsicherer Basis. *Zeitschrift für Fremdsprachenforschung* 20(1), S. 63–89.

Riis, U. & Francia, G. (2013). *Lärare, elever och spanska som modernt språk: Styrkor och svagheter – möjligheter och hot* [Lehrkräfte, Schüler und Spanisch als zweite Fremdsprache: Stärken und Schwächen – Möglichkeiten und Hindernisse]. Uppsala universitet: Fortbildningsavdelningen för skolans internationalisering.

Rinnert, C. & Kobayashi, H. (2001). Differing perceptions of EFL writing among readers in Japan. *The Modern Language Journal* 85, S. 189–209.

Robson, C. & McCartan, K. (2016). *Real world research: A resource for users of social research methods in applied settings* (4. Aufl.). Chichester: John Wiley & Sons Ltd.

Schneider, G.; Lenz, P. & Studer, T. (2009). *Fremdsprachen. Wissenschaftlicher Kurzbericht und Kompetenzmodell. Konsortium HarmoS Fremdsprachen*. Bern: EDK. <https://edudoc.educa.ch/static/web/arbeiten/harmos/L2_wissB_25_1_10_d.pdf> (Abgerufen: März 2020.)

Schneider, G.; Lenz, P.; Forster Vosicki, B.; in Zusammenarbeit mit Glaboniat, M.; Imig, A.; North, B.; Piccardo, E.; Schmidt, M. G. & Sugitani, M. (2017). *Gemeinsamer europäischer Referenzrahmen für Sprachen (GER) und Europäisches Sprachenportfolio (ESP). IDT – Bericht der SIG Arbeitsgruppe 3.1. Stand: 11.04.17.* <https://www.idt2017.ch/images/03_fachprogramm/02_sig/SIG_3-1-Bericht_GER_und_ESP_20170403-IDT2017.pdf> (Abgerufen: Oktober 2019.)

Shadish, W. R.; Cook, T. D. & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.

Shaw, S. D. & Weir, C. J. (2007). *Examining writing: Research and practice in assessing second language writing*. Cambridge: Cambridge University Press.

Skar, G. (2013). *Skrivbedömning och validitet: Fallstudier av skrivbedömning i svenskundervisning på gymnasiet* [Schreibbewertung und Validität: Fallstudien zur Bewertung von Textproduktionen im Schwedischunterricht am Gymnasium]. [Diss.] Stockholm: Institutionen för språkdidaktik, Stockholm universitet.

Skolinspektionen (2010). *Kontrollrättning av nationella prov i grundskolan och i gymnasieskolan* [Kontrollkorrektur nationaler Prüfungen in der Grund- und Gymnasialschule]. Stockholm: Skolinspektionen.

Skolinspektionen (2017). *Bedömningsprocessernas betydelse för likvärdigheten. Ombedömning av nationella prov 2016* [Die Bedeutung der Bewertungsprozesse für die Gleichwertigkeit. Kontrollkorrekturen nationaler Prüfungen 2016]. U2014/7535/GV. Stockholm: Skolinspektionen.

Skolinspektionen (2018). *Ombedömning av nationella prov 2017 – fortsatt stora skillnader* [Kontrollkorrekturen nationaler Prüfungen 2017 – weiterhin große Unterschiede]. 2017:342. Stockholm: Skolinspektionen.

Skolverket (2000). *Språk. Grundskola och gymnasieskola. Kursplaner, betygskriterier och kommentarer* [Sprache. Grundschule und Gymnasialschule. Rahmenpläne, Bewertungskriterien und Kommentare]. Stockholm: Skolverket.

Skolverket (2009). *Likvärdig betygssättning i gymnasieskolan? En analys av sambandet mellan nationella prov och kursbetyg* [Gleichwertige Benotung am Gymnasium? Eine Analyse der Korrelation zwischen zentralen Prüfungen und Kursnoten]. Rapport 338. Stockholm: Skolverket. <https://www.skolverket.se/download/18.6bfaca41169863e6a65884c/1553962030954/pdf2286.pdf> (Abgerufen: Mai 2020.)

Skolverket (2011a). *Ämnesplan för moderna språk*. [Schwedischer Lehrplan für *Moderna språk*]. Stockholm: Skolverket/Fritzes.

Skolverket (2011b). *Kommentarmaterial till kursplanen i moderna språk*. [Kommentarmaterial der schwedischen Lehrpläne für *Moderna språk*] (1. Aufl.). Stockholm: Skolverket/Fritzes.

Skolverket (2011c). *Kommunalt huvudmannaskap i praktiken* [Kommunale Trägerschaft in der Praxis]. Rapport 362. Stockholm: Skolverket.

Skolverket (2012). *Bedömning av språklig kompetens – En studie av samstämmigheten mellan Internationella språkstudien 2011 och svenska styrdokument. Skolverkets analyser 2012.* [Bewertung sprachlicher Kompetenz – Eine

Studie der Übereinstimmung zwischen der Internationalen Sprachstudie 2011 und schwedischen Curriculumdokumenten. Analysen der schwedischen Schulbehörde 2012]. Stockholm: Skolverket/Fritzes. <https://www.skolverket.se/download/18.6bfaca41169863e6a659e88/1553964475581/pdf2831.pdf> (Abgerufen: Oktober 2020.)

Skolverket (2017a). *Personalstatistik – åldersfördelning. Läsåret 2016/2017.* [Personalstatistik – Altersverteilung. Schuljahr 2016/2017]. <https://siris.skolverket.se/reports/rwservlet?cmdkey=common&geo=1&report=pers_alder&verksform=21&p_verksamhetsar=2016&hman=&lankod=&kommunkod=> (Abgerufen: April 2019.)

Skolverket (2017b). *PM. Elever i gymnasieskolan 2016/17* [PM. Schülerinnen und Schüler am Gymnasium 2016/17]. Dnr: 5.1.1-2016:883. <https://siris.skolverket.se/siris/sitevision_doc.getFile?p_id=542378> (Abgerufen: Januar 2019.)

Skolverket (2018a). *Redovisning av uppdrag om förändringar av nationella program i gymnasieskolan samt av förslag utifrån propositionen Ökade möjligheter till grundläggande behörighet på yrkesprogram och ett estetiskt ämne i alla nationella program* [Berichterstattung über Aufgaben zu Änderungen der nationalen Programme in der Gymnasialschule und Vorschläge auf der Grundlage des Gesetzesvorhabens Verbesserte Möglichkeiten für eine grundlegende Hochschulzugangsberechtigung für praktische Gymnasialausrichtungen und ein ästhetisches Fach in allen nationalen Ausrichtungen]. Dnr 2018:00570. <https://www.skolverket.se/getFile?file=3965> (Abgerufen: Oktober 2019.)

Skolverket (2018b). *Språksprånget* [Der Sprachsprung]. <https://www.skolverket.se/skolutveckling/kurser-och-utbildningar/sprakspranget---kompetensutveckling-for--larare-i-moderna-sprak> (Abgerufen: Mai 2021.)

Skolverket (2019a). *Analyser av likvärdig betygssättning mellan elevgrupper och skolor* [Analysen gleichwertiger Bewertung zwischen Schülergruppen und Schulen]. Rapport 475. Stockholm: Skolverket. <https://www.skolverket.se/getFile?file=4035> (Abgerufen: Mai 2019.)

Skolverket (2019b). *PM – Pedagogisk personal i skola och vuxenutbildning läsåret 2018/2019* [PM – Pädagogisches Personal in der Schule und Erwachsenenbildung Jahrgang 2018/2019]. Rapport 475. <https://www.skolverket.se/publikationer?id=4050> (Abgerufen: Mai 2020.)

Skolverket (2020a). *Analyser av likvärdig betygsättning i gymnasieskolan. Jämförelser mellan kursbetyg och kursprov* [Analysen einer gleichwertigen Benotung am Gymnasium. Vergleiche zwischen Fachnoten und Fachprüfungen]. Rapport 2020:3. <https://www.skolverket.se/download/18.1a8151cc170ae4599bce10/1585902805741/pdf6564.pdf> (Abgerufen: Oktober 2020.)

Skolverket (2020b). *Likvärdiga betyg och meritvärden. Ett kunskapsunderlag om modeller för att främja betygens och meritvärdernas likvärdighet* [Gleichwertige Noten und Meritwerte. Ein Bericht über Modelle, um die Gleichwertigkeit der Noten und der Meritwerte zu fördern]. Stockholm: Skolverket. <https://www.skolverket.se/download/18.614394bd171bb7d771b55c6/1606807236500/pdf7581.pdf> (Abgerufen: Mai 2021.)

Skolverket (2021a). *Ämnesplan i moderna språk*. [Schwedischer Lehrplan für *Moderna språk*]. Stockholm: Skolverket. <https://www.skolverket.se/download/18.7f8c152b177d982455e114e/1615884383182/%C3%84mnesplan_moderna%20spr%C3%A5k.pdf> (Abgerufen: April 2021.)

Skolverket (2021b). Bedömningsportalen [Das Bewertungsportal]. <https://bp.skolverket.se/web/bs_gy_modtys/start> (Abgerufen: Mai 2021.)

Skolverket (2021c). *Elever och skolenheter i grundskolan läsåret 2020/2021* [Schüler/Schülerinnen und Schuleinheiten in der Grundschule Jahrgang 2020/2021]. <https://www.skolverket.se/getFile?file=7920> (Abgerufen: Oktober 2021.)

Skolverket (2021d). *Genomföra och bedöma nationella prov i gymnasieskolan* [Durchführung und Bewertung nationaler Tests am Gymnasium]. Stockholm: Skolverket. <https://www.skolverket.se/undervisning/gymnasieskolan/nationella-prov-i-gymnasieskolan/genomfora-och-bedoma-prov-i-gymnasieskolan> (Abgerufen: April 2021.)

Skolverket (2021e). *Kommentarmaterial till ämnesplanerna i moderna språk och engelska. Gymnasieskolan och kommunal vuxenutbildning på gymnasial nivå* [Kommentarmaterial der schwedischen Lehrpläne für *Moderna språk* und Englisch. Das Gymnasium und kommunale Erwachsenenbildung auf gymnasialer Ebene]. Stockholm: Skolverket. <https://www.skolverket.se/publikationer?id=7842> (Abgerufen: April 2021.)

Song, B. & Caruso, I. (1996). Do English and ESL faculty differ in evaluating the essays of native English-speaking and ESL students? *Journal of Second Language Writing* 5(2), S. 163–182.

SOU 1942:11 (1942). *Betänkande med utredning och förslag angående betygssättning i folkskolan* [Bericht mit Ermittlung sowie Vorschläge zur Benotung in der schwedischen Volksschule]. Stockholm: Ecklesiastikdepartementet.

SOU 1948:27 (1948). *1946 års skolkommissions betänkande med förslag till riktlinjer för det svenska skolväsendets utveckling* [Der Bericht der Schulkommission aus dem Jahr 1946 mit Vorschlägen für Leitlinien zur Entwicklung des schwedischen Schulsystems]. Stockholm: Ecklesiastikdepartementet.

Stemler S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment Research & Evaluation* 9(4), S. 1–11.

Stemler, S. E. & Tsai, J. (2008). Best practices in interrater reliability. Three common approaches. In: J. W. Osborne (Hrsg.), *Best practices in quantitative methods*. Los Angeles: Sage, S. 29–49.

Stobart, G. (2003). The impact of assessment: intended and unintended consequences. *Assessment in Education: Principles, Policy & Practice* 10(2), S. 139–140.

Stobart, G. (2012). Validity in formative assessment. In: J. Gardner (Hrsg.), *Assessment and learning*. London: Sage, S. 233–242.

Sundquist. P. & Sylvén, L. K. (2014). Language-related computer use: Focus on young L2 English learners in Sweden. *Recall* 26(1), S. 3–20.

Sweedler-Brown, C. O. (1985). The influence of training and experience on holistic essay evaluations. *The English Journal* 74(5), S. 49–55.

Tengberg, M.; Borgström, E.; Lötmarker, L.; Sandlund, E.; Skar, G. B.; Sundqvist, P.; Walkert, M. & Wikberg, K. (2017). *Likvärdig bedömning av elevers språkförmågor: Preliminära resultat från ett ämnesdidaktiskt forskningsprojekt* [Gleichwertige Bewertung sprachlicher Kompetenzen von Schülerinnen und Schülern: Vorläufige Ergebnisse eines fachdidaktischen Forschungsprojekts]. Karlstad: Karlstad universitet.

Tholin, J. (2017). State control and governance of schooling and their effects on French, German, and Spanish learning in Swedish compulsory school, 1996–2011. *Scandinavian Journal of Educational research* 63(3), S. 317–332.

Tornberg, U. (2000). *Om språkundervisning i mellanrummet – och talet om „kommunikation" och „kultur" i kursplaner och läromedel från 1962 till 2000* [Zum Sprachunterricht im Zwischenraum – und die Rede über „Kommunikation" und „Kultur" in Lehrplänen und Lehrbüchern von 1962 zu 2000]. [Diss.] Uppsala: Acta Universitatis Upsaliensis.

Tornberg, U. (2015). *Språkdidaktik* (5. Aufl.). Malmö: Gleerups.

Torrance, H. (2007). Assessment *as* learning? How the use of explicit learning objectives, assessment criteria and feedback in post-secondary education and training can come to dominate learning. *Assessment in Education* 14(3), S. 281–294.

Toulmin, S. (1958). *The uses of argument*. London: Cambridge University Press.

Trace, J.; Janssen, G. & Meiner, V. (2017). Measuring the impact of rater negotiation in writing performance assessment. *Language Testing* 34(1), S. 3–22.

Tschirner, E. (2008). Vernünftige Erwartungen: Referenzrahmen, Kompetenzniveaus, Bildungsstandards. *Zeitschrift für Fremdsprachenforschung* 19(2), S. 187–208.

Tschirner, E. & Bärenfänger, O. (2012). *Assessing evidence of validity of assigning CEFR ratings to the ACTFL Oral Proficiency Interview (OPI) and the*

Oral Proficiency Interview by computer (OPIc). (Technical Report 2012-US-PUB-1.) Leipzig: Institute for Test Research and Development.

Tyllered, M. (2002). *Jämförelse Engelska C/steg 7 och Cambridge Certificate in Advanced English* [Vergleiche Englisch C/Stufe 7 und Cambridge Certificate in Advanced English]. [Unpubliziert.] Göteborg: Göteborgs universitet.

Utbildningsdepartementet (1993). *Högskoleförordningen (1993:100)* [Schwedische Verordnung des Hochschulwesens]. <https://www.riksdagen.se/sv/dokument-lagar/dokument/svensk-forfattningssamling/hogskoleforordning-1993100_sfs-1993-100> (Abgerufen: Oktober 2019.)

Utbildningsdepartementet (2010a). *Gymnasieförordningen (2010:2039)* [Schwedische Verordnung des Gymnasiums]. <https://www.riksdagen.se/sv/dokument-lagar/dokument/svensk-forfattningssamling/gymnasieforordning-20102039_sfs-2010-2039> (Abgerufen: Oktober 2019.)

Utbildningsdepartementet (2010b). *Skollagen (2010:801)* [Schwedisches Bildungsgesetz]. <https://www.riksdagen.se/sv/dokument-lagar/dokument/svensk-forfattningssamling/skollag-2010800_sfs-2010-800> (Abgerufen: Oktober 2019.)

Utbildningsdepartementet (2011). *Skolförordningen (2011:185)* [Schwedische Schulverordnung]. <https://www.riksdagen.se/sv/dokument-lagar/dokument/svensk-forfattningssamling/skolforordning-2011185_sfs-2011-185> (Abgerufen: Oktober 2019.)

Utbildningsdepartementet (2017). *Tillträde för nybörjare. Ett öppnare och enklare system för tillträde till högskoleutbildning* [Zugang für Anfänger. Ein offeneres und einfacheres System für den Zugang zu Hochschul- und Universitätsausbildung]. SOU 2017:020. <https://www.regeringen.se/49469a/contentassets/f170082619734b0082077bc95081f551/tilltrade-for-nyborjare--ett-oppnare-och-enklare-system-for-tilltrade-till-hogskoleutbildning-sou-201720.pdf> (Abgerufen: Oktober 2019.)

Utbildningsdepartementet (2018). *Redovisning av uppdrag om förslag på åtgärder i händelse av att meritpoängen avskaffas* [Bericht des Auftrages über Vorschläge von Maßnahmen im Fall, dass Meritpunkte abgeschafft werden] Dnr U2017/05023/BS, U2017/05037/S. <https://www.skolverket.se/getFile?file=3966> (Abgerufen: Oktober 2019.)

Vanbelle, S. (2016). A new interpretation of the weighted kappa coefficients. *Psychometrika* 81, S. 399–410.

Vaughan, C. (1991). Holistic assessment: what goes on in the rater's mind? In: L. Hamp-Lyons (Hrsg.), *Assessing second language writing in academic contexts*. Norwood, NJ: Ablex, S. 111–125.

Vetenskapsrådet (2002). *Forskningsetiska principer inom humanistisk-samhällsvetenskaplig forskning* [Empfehlungen zu forschungsethischen

Grundsätzen innerhalb Geistes- und Sozialwissenschaften]. Stockholm: Vetenskapsrådet.

Vlachos, J. (2019). Trust-Based Evaluation in a Market-Oriented School System. In: M. Dahlstedt & A. Fejes (Hrsg.), *Neoliberalism and Market Forces in Education: Lessons from Sweden*. London/New York: Routledge, S. 212–230.

Wahlström, N. (2016). *Läroplansteori och didaktik* [Lehrplantheorie und Didaktik] (2. Aufl.). Malmö: Gleerups.

Weigle S. C. (1994). Effects of training on raters of ESL compositions. *Language Testing* 11(2), S. 197–223.

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing* 15(2), S. 263–287.

Weigle, S. C. (2002). *Assessing Writing*. Cambridge. Cambridge University Press.

Weinert, F. E. (2001). Vergleichende Leistungsmessung in Schulen – eine umstrittene Selbstverständlichkeit. In: F. E. Weinert (Hrsg.): *Leistungsmessung in Schulen*. Weinheim und Basel: Beltz Verlag, S. 17–31.

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. Basingstoke: Palgrave Macmillan.

Wirtz, M. & Caspar, F. (2002). *Beurteilerübereinstimmung und Beurteilerreliabilität. Methoden zur Bestimmung und Verbesserung der Zuverlässigkeit von Einschätzungen mittels Kategoriensystemen und Ratingskalen*. Göttingen: Hogrefe-Verlag.

Wisniewski, K. (2010). Bewertervariabilität im Umgang mit GeR-Skalen. Ein- und Aussichten aus einem Sprachtestprojekt. *Deutsch als Fremdsprache* 3/2010, S. 143–149.

Wisniewski, K. (2014). *Die Validität der Skalen des gemeinsamen europäischen Referenzrahmens für Sprachen. Eine empirische Untersuchung der Flüssigkeits- und Wortschatzskalen des GeRS am Beispiel des Italienischen und des Deutschen*. [Diss.] Frankfurt am Main: Peter Lang.

Wu, R. (2011). *Validating second language reading examinations. Establishing the validity of the GEPT through alignment with the Common European Framework of Reference*. Cambridge: Cambridge University Press.

Xi, X. & Davis, L. (2016). Quality factors in language assessment. In: D. Tsagari, & J. Banerjee (Hrsg.), *Handbook of second language assessment*. Boston/Berlin: De Gruyter Mouton, S. 61–76.

Zhang, Y. & Elder, C. (2011). Judgements by oral proficiency of non-native and native English speaking teacher raters: competing or complementary constructs? *Language Testing* 28(1), S. 31–50.

Ziegler, N. & Kang, L. (2016). Mixed methods design. In: A. J. Moeller, J. W. Creswell & N. Saville (Hrsg.), *Second Language Assessment and Mixed Methods Research*. Cambridge: University Press, S. 51–83.

Anhang

Anhang 1: Zentrale Inhalte hinsichtlich Interaktion und Produktion in den schwedischen Bildungsstandards für *Moderna språk* bezüglich *Tyska 3*, *Tyska 4* und *Tyska 5*, hier im Original

Tyska 3	*Tyska 4*	*Tyska 5*
Instruktioner, berättelser och beskrivningar i sammanhängande tal och skrift. Diskussioner, samtal och skrivande för kontakt och kommunikation i olika situationer.	Instruktioner, berättelser och beskrivningar i sammanhängande tal och skrift. Samtal, diskussion och argumentation för kommunikation och kontakt i olika situationer.	Muntlig och skriftlig produktion och interaktion av olika slag, även i mer formella sammanhang, där eleverna instruerar, berättar, sammanfattar, förklarar, kommenterar, värderar, motiverar sina åsikter, diskuterar och argumenterar.
Strategier för att lösa språkliga problem, till exempel med hjälp av omformuleringar och förklaringar.	Strategier för att lösa språkliga problem, till exempel med hjälp av omformuleringar, frågor och förklaringar.	
Strategier för att bidra till och aktivt medverka i samtal, till exempel genom att ta initiativ till interaktion, lyssna aktivt och avsluta på ett artigt sätt.	Strategier för att bidra till och aktivt medverka i samtal, till exempel genom att ge bekräftelse, ställa följdfrågor och ta initiativ till nya frågeställningar eller ämnes-områden.	Strategier för att bidra till och aktivt medverka i diskussioner med anknytning till samhälls- och arbetslivet.
Språklig säkerhet när det gäller till exempel uttal, intonation, fasta språkliga uttryck och grammatiska strukturer, mot tydlighet, variation och anpassning till syfte, mottagare och situation.	Språklig säkerhet när det gäller till exempel uttal, intonation, fasta språkliga uttryck och satsbyggnad, mot tydlighet, variation och flyt.	
	Bearbetning av egna och andras muntliga och skriftliga framställningar för att variera, tydliggöra, precisera och anpassa dem till syfte, mottagare och situation.	Bearbetning av egna och andras muntliga och skriftliga fram-ställningar för att variera, tydliggöra och precisera dem samt för att skapa struktur och anpassa dem till syftet och situationen. I detta ingår användning av ord och fraser som tydliggör orsakssammanhang och tidsaspekter.

(Skolverket 2011a)

Anhang 2: Mindestanforderungen hinsichtlich Interaktion und Produktion in den schwedischen Bildungsstandards für *Moderna språk* bezüglich *Tyska 3*, *Tyska 4* und *Tyska 5*, hier im Original

Tyska 3	*Tyska 4*	*Tyska 5*
I muntliga och skriftliga framställningar av olika slag formulerar sig eleven **enkelt, begripligt och till viss del** sammanhängande. För att förtydliga och variera sin kommunikation bearbetar eleven, och gör **enkla** förbättringar av, egna framställningar.	I muntliga och skriftliga framställningar i olika genrer formulerar sig eleven **enkelt, begripligt** och **relativt** sammanhängande. För att förtydliga och variera sin kommunikation bearbetar eleven, och gör **enkla** förbättringar av, egna framställningar.	I muntliga och skriftliga framställningar i olika genrer formulerar sig eleven **relativt** varierat, **relativt** tydligt och **relativt** sammanhängande. Eleven formulerar sig även med **visst** flyt och **i någon mån anpassat** till syfte, mottagare och situation. Eleven bearbetar och gör **enkla** förbättringar av egna framställningar.
I muntlig och skriftlig interaktion uttrycker sig eleven **begripligt och enkelt**. Dessutom väljer och använder eleven **i huvudsak** fungerande strategier som **i viss mån** löser problem i och förbättrar interaktionen.	I muntlig och skriftlig interaktion i olika sammanhang uttrycker sig eleven **begripligt och enkelt** samt **i någon mån anpassat** till syfte, mottagare och situation. Dessutom väljer och använder eleven **i huvudsak** fungerande strategier som **i viss mån** löser problem i och förbättrar interaktionen.	I muntlig och skriftlig interaktion i olika, även mer formella, sammanhang uttrycker sig eleven tydligt och med **visst** flyt samt med **viss** anpassning till syfte, mottagare och situation. Dessutom väljer och använder eleven **i huvudsak** fungerande strategier som **i viss mån** löser problem i och förbättrar interaktionen.

(Skolverket 2011a)

Anhang 3: GER-Skala: Schriftliche Produktion allgemein

Schriftliche Produktion allgemein	
C2	Kann klare, flüssige, komplexe Texte in angemessenem und effektivem Stil schreiben, deren logische Struktur den Lesern das Auffinden der wesentlichen Punkte erleichtert.
C1	Kann klare, gut strukturierte Texte zu komplexen Themen verfassen und dabei die entscheidenden Punkte hervorheben, Standpunkte ausführlich darstellen und durch Unterpunkte oder geeignete Beispiele oder Begründungen stützen und den Text durch einen angemessenen Schluss abrunden.
B2	Kann klare, detaillierte Texte zu verschiedenen Themen aus seinem/ihrem Interessengebiet verfassen und dabei Informationen und Argumente aus verschiedenen Quellen zusammenführen und gegeneinander abwägen.
B1	Kann unkomplizierte, zusammenhängende Texte zu mehreren vertrauten Themen aus seinem/ihrem Interessengebiet verfassen, wobei einzelne kürzere Teile in linearer Abfolge verbunden werden.
A2	Kann eine Reihe einfacher Wendungen und Sätze schreiben und mit Konnektoren wie *und*, *aber* oder *weil* verbinden.
A1	Kann einfache, isolierte Wendungen und Sätze schreiben.

(Europarat 2001: 67)

Anhang 4: GER-Skala: Schriftliche Interaktion allgemein

Schriftliche Interaktion allgemein	
C2	Wie C1
C1	Kann sich klar und präzise ausdrücken und sich flexibel und effektiv auf die Adressaten beziehen.
B2	Kann Neuigkeiten und Standpunkte effektiv schriftlich ausdrücken und sich auf solche von anderen beziehen.
B1	Kann Informationen und Gedanken zu abstrakten wie konkreten Themen mitteilen, Informationen prüfen und einigermaßen präzise ein Problem erklären oder Fragen dazu stellen. Kann in persönlichen Briefen und Mitteilungen einfache Informationen von unmittelbarer Bedeutung geben oder erfragen und dabei deutlich machen, was er/sie für wichtig hält.
A2	Kann kurze, einfache, formelhafte Notizen machen, wenn es um unmittelbar notwendige Dinge geht.
A1	Kann schriftlich Informationen zur Person erfragen oder weitergeben.

(Europarat 2001: 86)

Anhang 5: GER-Skala: Wortschatzbeherrschung

Wortschatzbeherrschung	
C2	Durchgängig korrekte und angemessene Verwendung des Wortschatzes.
C1	Gelegentlich kleinere Schnitzer, aber keine größeren Fehler im Wortgebrauch.
B2	Die Genauigkeit in der Verwendung des Wortschatzes ist im Allgemeinen groß, obgleich einige Verwechslungen und falsche Wortwahl vorkommen, ohne jedoch die Kommunikation zu behindern.
B1	Zeigt eine gute Beherrschung des Grundwortschatzes, macht aber noch elementare Fehler, wenn es darum geht, komplexere Sachverhalte auszudrücken oder wenig vertraute Themen und Situationen zu bewältigen.
A2	Beherrscht einen begrenzten Wortschatz in Zusammenhang mit konkreten Alltagsbedürfnissen.
A1	Keine Deskriptoren verfügbar.

(Europarat 2001: 113)

Anhang 6: GER-Skala: Grammatische Korrektheit

Grammatische Korrektheit	
C2	Zeigt auch bei der Verwendung komplexer Sprachmittel eine durchgehende Beherrschung der Grammatik, selbst wenn die Aufmerksamkeit anderweitig beansprucht wird (z. B. durch vorausblickendes Planen oder Konzentration auf die Reaktion anderer).
C1	Kann beständig ein hohes Maß an grammatischer Korrektheit beibehalten; Fehler sind selten und fallen kaum auf.
B2	Gute Beherrschung der Grammatik; gelegentliche Ausrutscher oder nicht-systematische Fehler und kleinere Mängel im Satzbau können vorkommen, sind aber selten und können oft rückblickend korrigiert werden. Gute Beherrschung der Grammatik; macht keine Fehler, die zu Missverständnissen führen.
B1	Kann sich in vertrauten Situationen ausreichend korrekt verständigen; im Allgemeinen gute Beherrschung der grammatischen Strukturen trotz deutlicher Einflüsse der Muttersprache. Zwar kommen Fehler vor, aber es bleibt klar, was ausgedrückt werden soll. Kann ein Repertoire von häufig verwendeten Redefloskeln und von Wendungen, die an eher vorhersehbare Situationen gebunden sind, ausreichend korrekt verwenden.
A2	Kann einige einfache Strukturen korrekt verwenden, macht aber noch systematisch elementare Fehler, hat z. B. eine Tendenz, Zeitformen zu vermischen oder zu vergessen, die Subjekt-Verb-Kongruenz zu markieren; trotzdem wird in der Regel klar, was er/sie ausdrücken möchte.
A1	Zeigt nur eine begrenzte Beherrschung einiger weniger einfacher grammatischer Strukturen und Satzmuster in einem auswendig gelernten Repertoire.

(Europarat 2001: 114)

Anhang 7: GER-Skala: Beherrschung der Orthographie

Beherrschung der Orthographie	
C2	Die schriftlichen Texte sind frei von orthographischen Fehlern.
C1	Die Gestaltung, die Gliederung in Absätze und die Zeichensetzung sind konsistent und hilfreich. Die Rechtschreibung ist, abgesehen von gelegentlichem Verschreiben, richtig.
B2	Kann zusammenhängend und klar verständlich schreiben und dabei die üblichen Konventionen der Gestaltung und der Gliederung in Absätze einhalten. Rechtschreibung und Zeichensetzung sind hinreichend korrekt, können aber Einflüsse der Muttersprache zeigen.
B1	Kann zusammenhängend schreiben; die Texte sind durchgängig verständlich. Rechtschreibung, Zeichensetzung und Gestaltung sind exakt genug, so dass man sie meistens verstehen kann.
A2	Kann kurze Sätze über alltägliche Themen abschreiben – z. B. Wegbeschreibungen. Kann kurze Wörter aus seinem mündlichen Wortschatz „phonetisch“ einigermaßen akkurat schriftlich wiedergeben (benutzt dabei aber nicht notwendigerweise die übliche Rechtschreibung).
A1	Kann vertraute Wörter und kurze Redewendungen, z. B. einfache Schilder oder Anweisungen, Namen alltäglicher Gegenstände, Namen von Geschäften oder regelmäßig benutzte Wendungen abschreiben. Kann seine Adresse, seine Nationalität und andere Angaben zur Person buchstabieren.

(Europarat 2001: 118)

Anhang 8: GER-Skala: Kohärenz und Kohäsion

Kohärenz und Kohäsion	
C2	Kann einen gut gegliederten und zusammenhängenden Text erstellen und dabei eine Vielfalt an Mitteln für die Gliederung und Verknüpfung angemessen einsetzen.
C1	Kann klar, sehr fließend und gut strukturiert sprechen und zeigt, dass er/sie die Mittel der Gliederung sowie der inhaltlichen und sprachlichen Verknüpfung beherrscht.
B2	Kann verschiedene Verknüpfungswörter sinnvoll verwenden, um inhaltliche Beziehungen deutlich zu machen. Kann eine begrenzte Anzahl von Verknüpfungsmitteln verwenden, um seine/ihre Äußerungen zu einem klaren zusammenhängenden Text zu verbinden; längere Beiträge sind möglicherweise etwas sprunghaft.
B1	Kann eine Reihe kurzer und einfacher Einzelelemente zu einem linearen, zusammenhängenden Äußerung verbinden.
A2	Kann die häufigsten Konnektoren benutzen, um einfache Sätze miteinander zu verbinden, um eine Geschichte zu erzählen oder etwas in Form einer einfachen Aufzählung zu beschreiben. Kann Wortgruppen durch einfache Konnektoren wie *und*, *aber* und *weil* verknüpfen.
A1	Kann Wörter oder Wortgruppen durch sehr einfache Konnektoren wie *und* oder *dann* verbinden.

(Europarat 2001: 125)

GOETHE-ZERTIFIKAT B1	SCHREIBEN
ÜBUNGSSATZ j	KANDIDATENBLÄTTER

Kandidatenblätter

Schreiben
60 Minuten

Das Modul *Schreiben* besteht aus drei Teilen.

In den **Aufgaben 1** und **3**
schreibst du E-Mails.
In **Aufgabe 2**
schreibst du einen Diskussionsbeitrag.

Du kannst mit jeder Aufgabe beginnen.
Schreibe deine Texte auf die
Antwortbogen.

Bitte schreibe deutlich und
verwende keinen Bleistift.

Hilfsmittel wie z. B. Wörterbücher oder
Mobiltelefone sind nicht erlaubt.



Aufgabe 1 Arbeitszeit: 20 Minuten

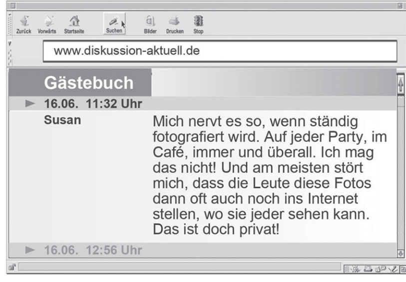
Du hast im Sommer ein Praktikum in einer Buchhandlung gemacht und möchtest einem Freund/einer Freundin davon erzählen.

- Beschreibe: Wie war das Praktikum in der Buchhandlung?
- Begründe: Was hat dir besonders gut gefallen?
- Mache einen Vorschlag für ein Treffen.

Schreibe eine E-Mail (circa 80 Wörter).
Schreibe etwas zu allen drei Punkten.
Achte auf den Textaufbau (Anrede, Einleitung, Reihenfolge der Inhaltspunkte, Schluss).

Aufgabe 2 Arbeitszeit: 25 Minuten

Du hast in einer Zeitschrift einen Artikel zum Thema „Private Fotos in sozialen Netzwerken" gelesen. Im Online-Forum der Zeitung findest du folgende Meinung:



Schreibe nun deine Meinung zum Thema (circa 80 Wörter).

Aufgabe 3 Arbeitszeit: 15 Minuten

Es ist Abend und du solltest deinem Deutschlehrer, Herrn Schmidt, bis heute eine Hausaufgabe per E-Mail zuschicken. Du hast sie aber noch nicht fertig gemacht.

Schreibe an Herrn Schmidt. Entschuldige dich *höflich* und begründe, warum du erst morgen fertig bist.

Schreibe eine E-Mail (circa 40 Wörter).
Vergiss nicht die Anrede und den Gruß am Schluss.

Seite 2

Anhang 9: Prüfungsteil: Schriftlicher Ausdruck des Goethe-Zertifikats B1

(Goethe-Institut 2017: 23–24) © Goethe-Institut, München. Für die Genehmigung des Abdrucks danke ich herzlich Frau Stefanie Dengler, Goethe-Institut, München.

Anhang 10: Tabellen zu den Hintergrundvariablen der teilnehmenden Bewertenden

Tab. 42: *Hintergrundvariablen der Gruppe der schwedischen Lehrkräfte (N = 18)*

Bewertende/r	*Schultyp*	*Altersspanne*	*Lehrererfahrung (Jahre)*
Lehrkraft 1	Kommunale Schule	< 60	20–29
Lehrkraft 2	kommunale Schule	40–49	10–19
Lehrkraft 3	freie Schule	< 60	30–39
Lehrkraft 4	kommunale Schule	40–49	10–19
Lehrkraft 5	kommunale Schule	30–39	1–9
Lehrkraft 6	kommunale Schule	40–49	1–9
Lehrkraft 7	freie Schule	30–39	10–19
Lehrkraft 8	freie Schule	50–59	20–29
Lehrkraft 9	kommunale Schule	50–59	30–39
Lehrkraft 10	kommunale Schule	40–49	20–29
Lehrkraft 11	kommunale Schule	50–59	30–39
Lehrkraft 12	kommunale Schule	50–59	30–39
Lehrkraft 13	kommunale Schule	50–59	30–39
Lehrkraft 14	freie Schule	40–49	20–29
Lehrkraft 15	kommunale Schule	< 60	30–39
Lehrkraft 16	kommunale Schule	50–59	10–19
Lehrkraft 17	kommunale Schule	50–59	20–29
Lehrkraft 18	kommunale Schule	< 60	30–39

Tab. 43: *Hintergrundvariablen der externen schwedischen Bewertenden*

Bewertende/r	*Schultyp*	*Altersspanne*	*Lehrererfahrung (Jahre)*
ext. schwed. Bewert. 1	kommunale Schule	< 60	< 40
ext. schwed. Bewert. 2	kommunale Schule	40–49	20–29

Tab. 44: *Hintergrundvariablen der GER-Bewertenden*

Bewertende/r	*Altersspanne*	*Bewertererfahrung (Jahre)*
GER-Bewert. 1	50–59	20–29
GER-Bewert. 2	30–39	1–9

Anhang 11: Beurteilungsfaktoren in den Anweisungen zum fakultativen Prüfungsmaterial in der zweiten Fremdsprache für die Bewertung schriftlicher Kompetenz

Innehåll (*Inhalt*)
• begriplighet och tydlighet (*Verständlichkeit und Deutlichkeit*)
• fyllighet och variation (*Fülle und Variation*)
– olika exempel och perspektiv (*verschiedene Beispiele und Perspektiven*)
• sammanhang och struktur (*Kohärenz und Struktur*)
• anpassning till syfte, mottagare och situation (*Anpassung an die Absicht von Geschriebenem sowie situations- und partneradäquat*)

Språk och uttrycksförmåga (*Sprache und Ausdrucksfähigkeit*)
• kommunikativa strategier, t.ex. omformuleringar, förklaringar och förtydliganden (*kommunikative Strategien, z. B. Umformulierungen, Erklärungen und Verdeutlichungen*)
• flyt och ledighet (*Flüssigkeit und Ungezwungenheit*)
• omfång, variation, tydlighet och säkerhet (*Umfang, Variation, Deutlichkeit und Sicherheit*)
– vokabulär, fraseologi och idiomatik (*Vokabular, Phraseologie und Idiomatik*)
– meningsbyggnad och textbindning (*Satzbau und Textbindung*)
– grammatiska strukturer (*grammatische Strukturen*)
– stavning och interpunktion (*Rechtschreibung und Interpunktion*)
• anpassning till syfte, mottagare och situation (*Anpassung an die Absicht von Geschriebenem sowie situations- und partneradäquat*)

(Skolverket 2021b)

ZERTIFIKAT B1	BEWERTUNGSKRITERIEN
MODELLSATZ	PRÜFERBLÄTTER

Bewertungskriterien Schreiben

			A	B	C	D	E
AUFGABE 1	**Erfüllung** *	Inhalt, Umfang, Sprachfunktionen (z. B. jemanden einladen, Vorschlag machen …)	Alle 3 Sprachfunktionen inhaltlich und umfänglich angemessen behandelt	2 Sprachfunktionen angemessen **oder** 1 angemessen und 2 teilweise	1 Sprachfunktion angemessen und 1 teilweise **oder** alle teilweise	1 Sprachfunktion angemessen **oder** teilweise	Textumfang weniger als 50 % der geforderten Wortanzahl **oder** Thema verfehlt
		Textsorte	durchgängig umgesetzt	erkennbar	ansatzweise erkennbar	kaum erkennbar	
		Register/ Soziokulturelle Angemessenheit	situations- und partneradäquat	noch weitgehend situations- und partneradäquat	ansatzweise situations- und partneradäquat	nicht mehr situations- und partneradäquat	
	Kohärenz	Textaufbau (z. B. Einleitung, Schluss …)	durchgängig und effektiv	überwiegend erkennbar	stellenweise erkennbar	kaum erkennbar	Text durchgängig unangemessen
		Verknüpfung von Sätzen, Satzteilen	angemessen	überwiegend angemessen	teilweise angemessen	kaum angemessen	
	Wortschatz	Spektrum	differenziert	überwiegend angemessen	teilweise angemessen **oder** begrenzt	kaum vorhanden	
		Beherrschung	vereinzelte Fehlgriffe beeinträchtigen das Verständnis nicht	mehrere Fehlgriffe beeinträchtigen das Verständnis nicht	mehrere Fehlgriffe beeinträchtigen das Verständnis teilweise	mehrere Fehlgriffe beeinträchtigen das Verständnis erheblich	
	Strukturen	Spektrum	differenziert	überwiegend angemessen	teilweise angemessen **oder** begrenzt	kaum vorhanden	
		Beherrschung (Morphologie, Syntax, Orthografie)	vereinzelte Fehlgriffe beeinträchtigen das Verständnis nicht	mehrere Fehlgriffe beeinträchtigen das Verständnis nicht	mehrere Fehlgriffe beeinträchtigen das Verständnis teilweise	mehrere Fehlgriffe beeinträchtigen das Verständnis erheblich	
AUFGABE 2	**Erfüllung** *	Inhalt, Umfang, Meinungsäußerung	Meinungsäußerung inhaltlich und umfänglich angemessen	überwiegend angemessen	teilweise angemessen	kaum angemessen	Wie Aufgabe 1
		Register/ Soziokulturelle Angemessenheit	situations- und partneradäquat	noch weitgehend situations- und partneradäquat	ansatzweise situations- und partneradäquat	nicht mehr situations- und partneradäquat	
	Kohärenz **Wortschatz** **Strukturen**		Wie Aufgabe 1				
AUFGABE 3	**Erfüllung** *	Mitteilung, Inhalt Register/ Soziokulturelle Angemessenheit	Mitteilung inhaltlich und soziokulturell angemessen	überwiegend angemessen	stellenweise angemessen	kaum angemessen	Wie Aufgabe 1
	Kohärenz **Wortschatz** **Strukturen**		Wie Aufgabe 1				

* Wird das Kriterium „Erfüllung“ mit E (0 Punkten) bewertet, ist die Punktzahl für diese Aufgabe insgesamt 0 Punkte.

Vs_11_070115

Seite 42

Anhang 12: Bewertungsraster für den Prüfungsteil *Schriftlicher Ausdruck* zum Goethe-Zertifikat B1

(Goethe-Institut 2017: 43) Für die Genehmigung des Abdrucks danke ich herzlich Frau Stefanie Dengler, Goethe-Institut, München.

Anhang 13: Beispiele von Textproduktionen schwedischer Schülerinnen und Schüler

Tyska 3: Textproduktion Hmlt2–3 (Lehrkraft: C; ext. schwed. Bewertender 1: F; ext. schwed. Bewertender 2: F; GER-Bewertung: nicht B1-Niveau; 30 Punkte)

1. Hallo Frida!

Im Sommer habe ich ein Praktikum in einer Buchhandlung. Die Praktikum war prima. Ich habe viel gelernt und die zeit auf die Buchhandlung war super spaß.

Die Praktikum war besonders gut gefallen für mich weil ich liebe Büch gelessen.

Ich habe viel zu erzählen für dich. Kannst du treffen mich auf die café am Donnerstag?

Deine NN!

2. Ich hasse wenn ständig fotografiert wird. Ich finde dass private Fotos will nicht auch die Internet stellen. Jeder kann sehen und ich mag das nicht.

3. Hallo Herrn Schmidt.

Ich weiß dass die Hausaufgabe solltest bin klar bis Heute. Aber ich bin nicht fertig jetz für ich habe meine großvater und großmutter besucht. Ich schreibt klar die hausaufgabe am morgen.

Tshüss!

NN

Tyska 4: Textproduktion Klju1–4 (Lehrkraft: C; ext. schwed. Bewertender 1: C; ext. schwed. Bewertender 2: C; GER-Bewertung: B1-Niveau; 88 Punkte)

1. Hallo!
Wie geht´s? Was hast du diese Sommer gemacht?
Ich habe ein Praktikum in einer Buchhandlung gemacht. Es war okay. Ich habe viele Menschen getroffen, und sehr viel über Litteratur gelernt. Aber, ich habe zuviel Bücher gesehen. Jetzt finde ich Bücher und besonders Büchhandlungen langweilig. Eine gute Sache hat es mir nur geben! Ich habe von viele Bücher gehört, dass ich sehr spannend findet. Als wir uns nächste mal treffen, muss ich dir ein par Bücher empfehlen.
Aber wann konntest du mir treffen? Was sagst du über nächsten Freitag?
Tschüss!
NN

2. Meine Meinung ist, dass wir müssen respektieren was privat ist. Wann jemand woll privat waren, müssen wir dass respektieren. Natürlich muss mann sich fragen, was die Leute in den Foto wirklich willst. Und weisst man nicht, könnt man sie fragen.

3. Guten Abend, Herr Schmidt!
Es tut mir leit, aber meine Hausaufgabe bis heute ist noch nicht fertig! Meine Mutter ist krank und ich muss ihr immer helfen. Morgen kommt mein Bruder, und wann er unsere Mutter hilft kann ich die Hausaufgabe fertig machen.
Vielen Dank!
NN

Tyska 5: Textproduktion Pnmj1–5 (Lehrkraft: F; ext. schwed. Bewertender 1: F; ext. schwed. Bewertender 2: E; GER-Bewertung: B1-Niveau; 73 Punkte)

1. Lieber Peter
Dieser Sommer habe ich mein Praktikum gemacht. Es war in einer alter Buchhandlung, der liegt bei der Fluss im Stadt. Ich war so froh, wenn ich in Buchhandlng am erst gewissen war, weil es riecht mit alte Buchen! Weil es so gut im Sommer gegangen hat, haben sie mich ein richtiger Arbeit geben. Ich arbeite jede Wochenende, so kannst du zu Buchhandlung kommen am Sonntag? Samstag funktioniert ganz gut auch!
A.d.b., NN

2. Ich finde dass es ist gut, wie viele Fotos sind ins Internet stellen, weil es dann gibt viel Material von unser Zeit, wann mann in ein Paar hunderen Jahren den studiert wollt. Ob die Fotos sind privat oder nicht ist mir egal – wie lang es gibt Fotos von die Alltagsleben des Menschen jetzt. Es ist doch nicht so gut, ob die Fotos zu perzönlich sind – dann können sie Probleme machen, wie ein Arbeitsinterview zerstören.

Undschuldigung, Herr Schmidt!
Ich habe nicht meine Hausaufgabe getun, weil mein Deutschbuch von meine Katze zerstört ist. Ich verstehe, dass es typisch ist, aber es ist ganz Wahr – Morgen kann ich das Buch zu Schule mitbringen, oder was von den überlebt. Viele Undschuldigungen, Herr Schmidt, bis Morgen will ich das schaffen! //NN

Nordeuropäische Arbeiten zur Literatur, Sprache und Kultur / Northern European Studies in Literature, Language and Culture

Herausgegeben von / Edited by Frank Thomas Grub

Band 1 Frank Thomas Grub (Hrsg.): Landeskunde Nord. Beiträge zur 1. Konferenz in Göteborg am 12. Mai 2012. 2013.

Band 2 Christine Becker / Frank Thomas Grub (Hrsg.): Perspektive Nord: Zu Theorie und Praxis einer modernen Didaktik der Landeskunde. Beiträge zur 2. Konferenz des Netzwerks *Landeskunde Nord* in Stockholm am 24./25. Januar 2014. 2015.

Band 3 Magnus P. Ängsal / Frank Thomas Grub (Hrsg.): Visionen und Illusionen. Beiträge zur 11. Arbeitstagung schwedischer Germanistinnen und Germanisten *Text im Kontext* in Göteborg am 4./5. April 2014. 2015.

Band 4 Erla Hallsteinsdóttir / Klaus Geyer / Katja Gorbahn / Jörg Kilian (Hrsg.): Perspektiven der Stereotypenforschung. 2016.

Band 5 Niclas Johansson: The Narcissus Theme from *Fin de Siècle* to Psychoanalysis. Crisis of the Modern Self. 2017.

Band 6 Klaus Geyer / Frank Thomas Grub (Hrsg.): Spektrum Nord: Vielfalt der Ziele, Inhalte und Methoden in der Landeskunde. Beiträge zur 3. Konferenz des Netzwerks *Landeskunde Nord* in Odense am 21./22. Januar 2016. 2017.

Band 7 Frank Thomas Grub / Dessislava Stoeva-Holm (Hrsg.): Emotionen: Beiträge zur 12. Arbeitstagung schwedischer Germanistinnen und Germanisten *Text im Kontext* in Visby am 15./16. April 2016. 2018.

Band 8 Heike Havermeier: Codeswitching als Mehrsprachigkeitspraxis in der universitären Kommunikation. Eine Untersuchung am Beispiel von Germanisten in Schweden. 2020.

Band 9 Frank Thomas Grub / Maris Saagpakk (Hrsg.): Brückenschläge Nord: Landeskunde an der Schnittstelle von Schule und Universität. Beiträge zur 4. Konferenz des Netzwerks *Landeskunde Nord* in Tallinn am 26./27. Januar 2018. 2020.

Band 10 Klaus Geyer / Anke Heier / Mette Skovgaard Andersen (Hrsg.): Tysk(a) – saksa – vācu – vokiečių – þýska 2020. Teil 1: Deutsche Sprachwissenschaft und Sprachdidaktik. Ausgewählte Beiträge zum *XI. Nordisch-Baltischen Germanistentreffen* in Kopenhagen vom 26.–29. Juni 2018. 2021.

Band 11 Mirjam Gebauer / Maris Saagpakk (Hrsg.): Tysk(a) – saksa – vācu – vokiečių – þýska 2020. Teil 2: Germanistische Literatur- und Kulturwissenschaft. Ausgewählte Beiträge zum *XI. Nordisch-Baltischen Germanistentreffen* in Kopenhagen vom 26.-29. Juni 2018. 2021.

Band 12 Janina Gesche: Stockholmer literarische Entscheidungen. Zu den Ausleseprozessen bei der Vergabe des Nobelpreises für Literatur am Beispiel deutschsprachiger Kandidaten – von Theodor Mommsen bis Hermann Hesse. 2021.

Band 13 Hanna Henryson: Gentrifiktionen. Zur Gentrifizierung in deutschsprachigen Berlin-Romanen nach 2000. 2021.

Band 14 Ingemar Haag: Stockholmer literarische Entscheidungen. Negotiating Selfhood in Self-Representational Works by Goethe, Sand, and Nietzsche. 2021.

Band 15 Helga Müllneritsch: The Austrian Manuscript Cookery Book in the Long Eighteenth Century. Studies of Form and Function. 2022.

Band 16 Agneta Hauber: Melitta Urbancic: Lyrik am Rand der Welt. Exil und Integration in Island. 2022.

Band 17 Maria Håkansson Ramberg: Validität und schriftliche Sprachkompetenz. 2023.

www.peterlang.com